CURRENT ISSUES IN COMPUTING AND PHILOSOPHY

VISIT…

# Frontiers in Artificial Intelligence and Applications

FAIA covers all aspects of theoretical and applied artificial intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes. The FAIA series contains several sub-series, including “Information Modelling and Knowledge Bases” and “Knowledge-Based Intelligent Engineering Systems”. It also includes the biennial ECAI, the European Conference on Artificial Intelligence, proceedings volumes, and other ECCAI – the European Coordinating Committee on Artificial Intelligence – sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.



## Volume 175

*Recently published in this series*

Vol. 174. S. Borgo and L. Lesmo (Eds.), Formal Ontologies Meet Industry
Vol. 173. A. Holst et al. (Eds.), Tenth Scandinavian Conference on Artificial Intelligence – SCAI 2008
Vol. 172. Ph. Besnard et al. (Eds.), Computational Models of Argument – Proceedings of COMMA 2008
Vol. 171. P. Wang et al. (Eds.), Artificial General Intelligence 2008 – Proceedings of the First AGI Conference
Vol. 170. J.D. Velásquez and V. Palade, Adaptive Web Sites – A Knowledge Extraction from Web Data Approach
Vol. 169. C. Branki et al. (Eds.), Techniques and Applications for Mobile Commerce – Proceedings of TAMoCo 2008
Vol. 168. C. Riggelsen, Approximation Methods for Efficient Learning of Bayesian Networks
Vol. 167. P. Buitelaar and P. Cimiano (Eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge
Vol. 166. H. Jaakkola, Y. Kiyoki and T. Tokuda (Eds.), Information Modelling and Knowledge Bases XIX
Vol. 165. A.R. Lodder and L. Mommers (Eds.), Legal Knowledge and Information Systems – JURIX 2007: The Twentieth Annual Conference
Vol. 164. J.C. Augusto and D. Shapiro (Eds.), Advances in Ambient Intelligence
Vol. 163. C. Angulo and L. Godo (Eds.), Artificial Intelligence Research and Development
Vol. 162. T. Hirashima et al. (Eds.), Supporting Learning Flow Through Integrative Technologies

ISSN 0922-6389

# Current Issues in Computing and Philosophy

Edited by

Adam Briggle

Katinka Waelbers

and

Philip A.E. Brey

*Department of Philosophy, University of Twente, The Netherlands*

IOS Press

Amsterdam • Berlin • Oxford • Tokyo • Washington, DC

ISBN 978-1-58603-876-2
Library of Congress Control Number: 2008928220

*Publisher*
IOS Press
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

*Distributor in the UK and Ireland*
Gazelle Books Services Ltd.
White Cross Mills
Hightown
Lancaster LA1 4XS
United Kingdom
fax: +44 1524 63232
e-mail: sales@gazellebooks.co.uk

*Distributor in the USA and Canada*
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
e-mail: iosbooks@iospress.com



PRINTED IN THE NETHERLANDS

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Introduction

This volume collects eighteen essays presented at the fifth annual European Conference on Computing and Philosophy (ECAP) held June 21–23, 2007, at the University of Twente, the Netherlands. It represents some of the best of the more than eighty papers delivered at the conference. The theme of ECAP 2007 was the multi-faceted "computational turn" that is occurring through the interaction of the disciplines of philosophy and computing [1]. It was organized under the supervision of the International Association of Computing and Philosophy (IACAP). IACAP promotes scholarly dialogue and research on all aspects of the computational and informational turn, and on the use of information and communication technologies in the service of philosophy.

There are good reasons for both computer and information scientists and philosophers to do research at the intersection of computing and philosophy. In computer and information science, there are significant conceptual and methodological questions that require reflection and analysis. What, for example, is information? What are good methods in computer science, and can computer science be a genuine science? Is true artificial intelligence possible? These are questions asked by computer scientists and philosophers alike. Moreover, digital and information and communication technologies have had tremendous impact on society, which raises further philosophical questions. How, for example, are digital technologies changing our conception of reality, or of knowledge and information? How is the Internet changing human life, and is it creating a digital existence for us next to a physical existence? What are the ethical questions that both computer scientists and users of information technology are facing?

ECAP 2007, with over a hundred participants, has left us, the editors, with a sense that the multidisciplinary constellation of computing and philosophy is a vibrant and dynamic field, which has its best days still ahead of it. The manifold explorations at the intersection of computing and philosophy are yielding insights of both intrinsic interest and societal value. Yet, this multidisciplinary endeavor also presents a challenge. Like other such endeavors, it must continuously work to ensure that its diversity of perspectives and methods proves a source of strength and collaboration rather than a source of instability and disintegration. In short, we must always strive to communicate across boundaries.

It is our hope that the present volume facilitates this task. This raises a more specific challenge facing us as editors. Regrettably, we simply could not include all of the papers presented at ECAP. In making our selection, our guiding ambition was to create a snapshot of the field that would be of interest to an audience as diverse as ECAP itself. With that goal in mind, we have compiled top quality, accessible essays representing many of the most important areas of inquiry in the field.

In organizing the essays, we have taken a topical approach. The first three contributions explore the phenomenon of virtual worlds. The essays discuss ethical, anthropological, and ontological issues regarding such worlds and the avatars that inhabit them. The next four chapters focus on robots and artificial agents. They cover issues regarding human-robot interaction, agency in robots, and the social and ethical aspects of robotics, including military applications. The next group of chapters discusses the

relation between human mentality and information processing in computers. The essays consider the nature of representations in digital systems, the relations between data, information and knowledge, the relationships between computers and their users, and the nature of synthetic emotions. The final section covers a broad range of issues at the interface of computers and society. The cases discussed here include the educational potential of an intelligent tutoring system and a novel computer programming language, the integration of ethical principles into software design, the underrepresentation of women in computer science studies, and the way Internet users assess the trustworthiness of the information they encounter online.

We would like to thank IACAP and the ECAP steering committee for giving us the opportunity to organize the conference in 2007 and for helping us in publishing this volume. For guaranteeing high scientific quality, all contributions of this volume have been read by several referees. Here, we would like to thank them for their time and their vital contribution to this volume. Special thanks go to our Master student Maurice Liebregt who put a substantial amount of time and effort into the layout of the chapters and who very patiently incorporated all last minute corrections.

[1] L. Burkholder (ed). Philosophy and the Computer. Boulder, San Francisco, Oxford: Westview Press 1992.

# Contents

Introduction v

**Part I. Me, My Avatar, and I: Exploring Virtual Worlds**

Meta Ethics for the Metaverse: The Ethics of Virtual Worlds 3
*Edward H. Spence*

On the Ecological/Representational Structure of Virtual Environments 13
*Omar Rosas*

The Dynamic Representation of Reality and of Ourself between Real and Virtual Worlds 24
*Lukasz Piwek*

**Part II. Living with the Golem: Robots and Autonomous Agents**

Can a Robot Intentionally Conduct Mutual Communication with Human Beings? 35
*Kayoko Ishii*

On the Ethical Quandaries of a Practicing Roboticist: A First-Hand Look 45
*Ronald C. Arkin*

How Just Could a Robot War Be? 50
*Peter M. Asaro*

Limits to the Autonomy of Agents 65
*Merel Noorman*

**Part III. Mind and World: Knowing, Thinking, and Representing**

Formalising the 'No Information without Data-Representation' Principle 79
*Patrick Allo*

The Computer as Cognitive Artifact and Simulator of Worlds 91
*Philip Brey*

The Panic Room: On Synthetic Emotions 103
*Jordi Vallverdú and David Casacuberta*

Representation in Digital Systems 116
*Vincent C. Müller*

Information, Knowledge and Confirmation Holism 122
*Steve McKinlay*

Phenomenal Consciousness: Sensorimotor Contingencies and the Constitution of Objects 133
*Bastian Fischer and Daniel Weiller*

**Part IV. Computing in Society: Designing, Learning, and Searching**

Towards an Intelligent Tutoring System for Propositional Proof Construction 145
*Marvin Croy, Tiffany Barnes and John Stamper*

Toward Aligning Computer Programming with Clear Thinking via the Reason Programming Language 156
*Selmer Bringsjord and Jinrong Li*

Ethics and the Practice of Software Design 171
*Matteo Turilli*

How to Explain the Underrepresentation of Women in Computer Science Studies 184
*Margit Pohl and Monika Lanzenberger*

How the Web Is Changing the Way We Trust 194
*Dario Taraborelli*

Author Index 205

# Part I

# Me, My Avatar, and I: Exploring Virtual Worlds

This page intentionally left blank

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Meta Ethics for the Metaverse: The Ethics of Virtual Worlds

Edward H. SPENCE
*Department of Philosophy*
*University of Twente, The Netherlands*

**Abstract.** After a brief introduction that sets out the overall argument of the paper in summary, the second part of the paper will offer a meta-ethical framework based on the moral theory of Alan Gewirth, necessary for determining what, if any, ought to be the ethics that guide the conduct of people participating in virtual worlds in their roles as designers, administrators and players or avatars. As virtual worlds and the World Wide Web generally, are global in scope, reach and use, Gewirth's theory, which offers a supreme principle of morality, the Principle of Generic Consistency (PGC) that establishes universal rights for all persons always and everywhere, is particularly suitable for this task. The paper will show that persons both in the real world and in virtual worlds have rights to freedom and wellbeing. Strictly with regard to *agency* those rights are merely prima facie but with regard to *personhood* framed around the notion of *self-respect* those rights are absolute. The third and final part of the paper will examine in more practical detail and application why and how designers, administrators and avatars of virtual worlds are rationally committed on the basis of their own intrinsic purposive agency to ethical norms of conduct that require the universal respect of the rights of freedom and well-being of all agents, including their own. Using Alan Gewirth's argument for the Principle of Generic Consistency (*Reason and Morality*, 1978) and my expanded argument for the PGC in my *Ethics Within Reason: A Neo-Gewirthian Approach* (2006), the paper will specifically seek to demonstrate that insofar as avatars can be viewed as virtual representations or modes of presentations of real people (at least with regard to some virtual worlds in which the virtual agency of the avatar can be considered an extension of the agency of the person instantiating the avatar in the real world) and thus can and must be perceived as *virtual purposive agents*, then they have moral rights and obligations similar to those of their real counterparts. Finally, the paper will show how the rules of virtual worlds as instantiated by the designers' code and the administrators' end-user license agreement (EULA), must always be consistent with and comply with the requirements of universal morality as established on the basis of the PGC. When the two come into conflict, the PGC, as the supreme principle of morality, is always overriding.

**Keywords.** Ethics, virtual worlds, Gewirth

## Introduction

If virtual worlds are merely "virtual" and thus not real, why should we care about what happens in those worlds, let alone care about what the ethics of virtual worlds are or ought to be? A simple and straightforward answer to this question is that insofar as ethics concerns the inter-relations that people have with one another and insofar as such inter-relations can and do take place within the boundaries of virtual worlds, then clearly ethics is relevant to virtual worlds. If people of their own free will and informed consent decide to engage with each other in mercantile or any other types of social transactions within the boundaries of a virtual world, as they often do for example in the virtual world of *Second Life*, then those transactions ought to be bound by similar ethical standards as those applicable in the real world.

Banks such as ABN Amro, computer companies such as IBM, Universities such as the University of Oxford, famous fashion houses such Armani, to name but a few, all have business venues within Second Life. Thus, insofar as the *virtual* transactions that take place between people and these corporate institutions within the boundaries of virtual worlds involve inter-relationships between real persons, more precisely their virtual representations, then those virtual transactions, like transactions in the real world, ought to be bound by ethical norms and standards. By parity of argument, all inter-relationships formed between persons or more precisely their virtual representations in virtual worlds, ought to be bound by ethical standards, just as they are in the real world. So for example, if an inter-transaction between two individuals that involves one individual (A) cheating another individual (B) in a way that results in an injustice to individual (B), that inter-transaction is no less unethical with regard to the specific act of cheating merely because the cheating took place within the confines of a virtual world. The same general consideration applies to other forms of unethical acts that take place in virtual worlds and result in some form of harm or injustice caused by one individual or group of individuals to another individual or group of individuals.

## 1. Overall Argument of the Paper in Summary

For the purpose of this paper I shall define virtual worlds as "persistent, computer-mediated environments in which a plurality of players can interact with the world and each other" [1]. *Second Life and EverQuest,* among others, are such virtual worlds. Do players of virtual worlds and their avatar representatives in those worlds have rights? In this paper I shall argue that they do. Taking my cue from Ralph Koster's "declaration of the rights of avatars" [2] I shall base my claim on Alan Gewirth's argument for the Principle of Generic Consistency (PGC) which demonstrates that all purposive agents have *generic rights* to freedom and wellbeing [3]. I shall adopt that argument to demonstrate that insofar as avatars can be viewed as the virtual representations of the persons that instantiate them in the real world, and these persons have goals or purposes which they seek to fulfil within the environments of virtual worlds through their avatars, then they and by extension their avatars are acting purposively and as such, have rights to freedom and wellbeing. These rights, being only prima facie implies that they cannot be used by agents or their avatars to violate the legitimate rights of other purposive agents or those of their avatars.

I argue, moreover, that although only prima facie with regard to agency, rights to freedom and wellbeing become absolute and inalienable when they refer to the dignity of persons [4]. Thus although an agent's avatar could be justified in violating (I shall refer to a justified violation of rights as an *infringement*) the generic rights of another agent's avatar by killing them in a duel or combat let us say, especially when the code of the virtual world and the end-user license agreement (EULA) allows for that to happen or at least does not disallow it, no agent is ever justified in violating the rights of another agent by undermining their dignity or self-respect by some widely recognized act of degradation such as racial vilification, or rape, for example. Insofar as rape can take place in a virtual world (in some specified sense that renders it at least equivalent to real rape with regard to the degradation suffered by the victim), the victim of such virtual rape may be psychologically and emotionally harmed by being made to feel degraded.

I argue that in general, the code of a virtual world (VW) together with its EULA may allow or at least not disallow virtual "crimes" such as theft or killing that infringe an avatars' rights, provided those virtual crimes are in keeping with the accepted rules of the game played in that virtual world in accordance with the VW's code and EULA. However, a code or EULA should never allow and must always disallow virtual crimes in a virtual world or other acts that degrade or can potentially degrade the dignity of an avatar's person, such as virtual rape for example, even if this is intended as merely part of a game within a virtual world. With regard to absolute rights that a person has to one's dignity, morality both in the real world and within a virtual world is always permeable and porous, although it may be less so in the case of crimes that although may infringe an avatar's generic rights in a virtual world, such as theft and killing, do not violate but maintain respect for the avatar's absolute rights to their dignity as a person.

Thus contrary to Edward Castronova [5] I argue that morally speaking, virtual worlds can never be "closed" with regard to what affects or could potentially affect the personal dignity of avatars and their persons (the persons who instantiate them in the real world). There is, in other words, no moral *magic circle* separating virtual worlds from the real world, specifically with regard to the absolute rights that agents have to their dignity as persons, both within and without the boundaries of virtual worlds. With regard to morality but not always with regard to the law, there is an ethical continuum that runs between virtual worlds and the real world. An insult that causes offence can be just as hurtful within a virtual world as it can in a real world. An insult can potentially be morally *real* in both worlds.

I am thus in agreement with Jack Balkin that "the boundaries between the game space and real space are permeable" [6]. However, adopting a middle position between Edward Castronova and Jack Balkin, I claim that virtual worlds can under appropriate *interration laws* [5] as instantiated by the virtual world's code and EULA, allow for some closure that renders avatars that steal from or kill other avatars, immune from both moral culpability and legal sanction, especially if such actions are accepted by the avatars themselves as being part of the game space and game plan of the VW.

In exercising their right to free association and the right to freedom to play in virtual worlds that in their view enhances both their sense of freedom and wellbeing, avatars may choose to waive their prima facie rights to freedom and wellbeing that in the real world would preclude others from stealing from them and even killing them. Within a virtual world such actions may be permitted as being part of the "game" and thus morally acceptable within the *role morality* (see section 2) of the VW. However,

nothing within a virtual world that in some way degrades the personal dignity of an avatar may be permitted, even when it is in accordance with the role morality of the VW as instantiated by the VW's code and EULA.

An overriding proviso that needs to be emphasized, therefore, is that the code and EULA of a virtual world must always be consistent with and not contravene the requirements of the Principle of Generic Consistency (PGC), especially as they apply to respect for the absolute rights to freedom and wellbeing that all avatars have by virtue of their dignity as persons.

## 2. The Meta-Ethical Framework Informing the Argument

### *2.1. The rights of agents: Alan Gewirth's argument for the Principle of Generic Consistency*[1]

Due to constraints of space, I am unable to present a full exposition and detailed justification for Alan Gewirth's argument for the Principle of Generic Consistency (PGC) in this paper. I have done so in my book *Ethics Within Reason* [4]. As such, I will only provide a summarized exposition and Gewirth's argument for the PGC in outline only.

Gewirth's main thesis is that every rational agent, in virtue of engaging in action, is logically committed to accept a supreme moral principle, the Principle of Generic Consistency. The basis of his thesis is found in his doctrine that action has a normative structure, and because of this structure every rational agent, just in virtue of being an agent, is committed to certain necessary prudential and moral constraints.

Gewirth undertakes to prove his claim that every agent, *qua* agent, is committed to certain prudential and moral constraints in virtue of the normative structure of action in three main stages. First, he undertakes to show that by virtue of engaging in voluntary and purposive action, every agent makes certain implicitly evaluative judgments about the goodness of their purposes, and hence about the necessary goodness of their freedom and wellbeing, which are the necessary conditions for the fulfilment of their purposes. Secondly, he undertakes to show that by virtue of the necessary goodness which an agent attaches to his freedom and wellbeing, the agent implicitly claims that they have rights to these. These natural rights being, at this stage of the argument, only self-regarding are merely *prudential rights*. Thirdly, Gewirth undertakes to show that every agent must claim these rights in virtue of the sufficient reason that they are a *prospective purposive agent* (PPA) who have purposes they want to fulfil. Furthermore, every agent must accept that, since they have rights to their freedom and wellbeing for the sufficient reason that they are a PPA, they are logically committed, on pain of self-contradiction, to also accept the rational generalization that all PPAs have rights to freedom and wellbeing [3]. At this stage of the argument these rights being also other-regarding, now become *moral rights*. The conclusion of Gewirth's argument for the PGC is in fact a generalized statement for the PGC, namely, that all PPAs have rights to their freedom and wellbeing.

### *2.2. The absolute right to dignity*

#### *2.2.1. A reconstruction of Gewirth's argument for the PGC*

My reconstruction of Gewirth's argument for the PGC around the notion of self-respect and dignity (both personal and communal) in my book *Ethics Within Reason* [4] shows that an agent must not only claim rights to their freedom and wellbeing on the basis that these are the necessary conditions for all their purposive actions, but they must also claim rights to their freedom and wellbeing because these are the essential and fundamental constituents of their self-respect or dignity[2] (both personal and communal). In sum, an agent must consider that they have rights to their freedom and wellbeing not only because they are the sort of being who engages in voluntary and purposive action—that is to say, a being who is a *PPA*—but also because they are the sort of being who needs self-respect—that is to say, a being who is a *person*.

#### *2.2.2. The concept of absolute rights*

We can now see that to some degree at least, a person has the generic rights in virtue of being a person irrespective of what he does or omits to do as an agent. For every person, no matter what they do or fail to do, need their self-respect. Because all persons need their self-respect equally in virtue of being persons, each person will need a certain degree of freedom and wellbeing, especially the latter, in order to preserve and maintain a minimal degree of self-respect so as to preserve and maintain their personhood. Thus, a criminal needs their self-respect as much as a law-abiding citizen. In this sense, they must both have sufficient freedom and wellbeing to allow them to preserve and maintain their self-respect[3]. To the extent that a person has a right to have enough freedom and wellbeing in order to maintain their self-respect, that right is absolute. The right to minimal freedom and wellbeing, sufficient for a person to preserve and maintain their self-respect, cannot be removed without at the same time removing the very conditions necessary for an agent's personhood.

According to Gewirth,

> a right is absolute when it cannot be overridden in any circumstances, so that it can never be justifiably infringed and it must be fulfilled without any exceptions [7].

#### *2.2.3. Role morality and universal public morality [8]*

Every practice, profession or institution has its own internal *role morality***;** a morality determined by the specific overarching role of a particular practice, profession, or institution. Thus, the role of a police officer is to uphold law and order and to provide assistance in the criminal and judicial process; the role of a journalist is to inform the public truthfully and fairly on matters of public interest. The role morality of a particular practice, profession or institution sets in turn its own internal rules and codes of conduct for the ethical regulation of that practice, profession, or institution. Thus, typically, the code of ethics for a particular profession, industry or institution, would reflect and be constitutive of the role morality of that profession, industry or institution.

In contrast to role morality, I shall refer collectively to the moral requirement of equal respect of the rights to freedom and wellbeing of all purposive agents, in their

dual capacity as agents and persons, established on the basis of the argument for the Principle of Generic Consistency, as *universal public morality*.

Sometimes the role morality of a particular institution or profession may come into conflict with universal public morality. For example, a journalist might in the hope of getting a scoop for his media organisation violate a person's right to privacy. When that happens, universal public morality will always take precedence over role morality for the simple reason that universal public morality being foundational is overriding as it applies equally to everyone irrespective of the particular personal, professional, and institutional interests or other commitments, including those required by the role morality of a particular institution or profession.

Ultimately, the role morality of every institution and profession is answerable to the principles and hence the requirements of universal public morality because it is universal public morality that provides the foundational justification of any particular role morality. For it would be self-defeating to allow role morality to override the very principles of universal public morality that provide the initial and foundational moral justification of institutional or professional role morality.

## 3. The Meta-Ethical Framework Applied to the Ethics of Virtual Worlds

I have provided in section (2.2) a meta-ethical framework, which includes a summarized account of the essential features of Gewirth's argument for the Principle of Generic Consistency as the supreme and universal principle of morality, the reconstruction of the argument for the PGC around the concept of dignity, as well as a distinction between universal public morality and role morality. With the meta-ethical framework now in place, I can now outline the significance and consequences of the application of that meta-ethical framework for the ethics of virtual worlds. This is required for providing a rational foundation and justification for the introductory claims I made earlier in my overall argument for this paper in section (1).

### *3.1. The rights of virtual agents*

In this section I will attempt to demonstrate that insofar as avatars can be viewed as virtual representations, or virtual extensions, or modes of presentations of purposive agents in the real world, avatars, who just like their counterparts in the real world act and think purposively in virtual worlds, have rights to freedom and wellbeing. They have these rights on the basis of their virtual purposive agency. Strictly as agents who engage in purposive action they have these rights prima facie. However, as persons and specifically with regard to their self-respect or dignity, they have those rights absolutely. These rights of course extend to the designers and administrators of the virtual worlds, since they too are purposive agents and thus entitled to the same generic rights to freedom and wellbeing as the players and their avatars.

Insofar as purposive agency is the sufficient condition for having moral rights and insofar as the virtual agency of avatars can be viewed as an extension of the agency of the persons who instantiate them in the real world, it makes no difference, in principle at least, whether the purposive agency is that of real persons or that of avatars. Virtual purposive agency is as sufficient for establishing the rights of avatars as it is for

establishing the rights of real persons. In a sense, avatars are real persons that just happen to inhabit a virtual environment.

Given that avatars, with regard to their personal and communal dignity, hold their generic rights absolutely, codes and EULAs of virtual worlds are never morally justified in allowing the violation of those rights. Insofar as virtual rape can take place in a virtual world (I leave the matter open whether it can or it cannot) that would constitute a violation of an avatar's absolute rights and would thus be morally objectionable even if rape was somehow allowed by the code or EULA of the virtual world in question.

However, with regard to the avatars' prima facie rights to freedom and wellbeing, avatars may choose to waive their generic rights not to be killed or have their virtual property stolen, if the code and EULA of a particular virtual world allow such activities as part of a game.

### *3.1.1. Objection: only real agents can have rights*

An objection that could be raised against my argument so far is as follows: Can avatars be viewed in any meaningful way as *purposive agents (PAs)*? Only players are PAs. Avatars as mere virtual extensions or modes of presentations of their players, cannot be PAs and hence do not have rights.

To answer the above objection we must first briefly explore the question of who suffers moral harm. Is it the player or their avatar? This in turn leads to another question. What is the identity relationship of player and avatar? Is an avatar merely a player's different mode of presentation[4]? Does the avatar constitute a different identity-sense (or mode of presentation) but same identity-reference to that of player, suggested perhaps by Gottlob Freger's distinction between sense and reference? To summarise: When moral harm occurs, whose rights are infringed or violated, a player's or those of their avatar? If the player's, then avatars don't have rights and hence cannot suffer moral harm by violation of those rights.

### *3.1.2. Response to objection: room for rights*

Rights require a claimant and a respondent and a *world-space* in which rights can be actioned. Let us consider that a moral harm in a virtual world (VW) is a harm done by avatar A/Player X (AX -respondent) to avatar B/Player Y (BY-claimant). Lets us also assume that the respective identities of players Y and X are opaque: Y and X unknown to each other in both the real world (RW) and the virtual world (VW). In most cases, that would be the default position, for it is only optional that avatars choose (or not) to reveal their player-identities to each other. So in our example, players X and Y are in RW whereas their avatars A and B are in VW. Due to the opacity of world spaces (OWS), Claimant-BY can only seek moral redress against Respondent-AX for the moral harm done to him in VW but not RW. Hence, Avatar B as Player Y's *virtual representative (claimant)* can seek moral redress from Avatar A as Player X's virtual representative (respondent) for the moral harm done to them in VW since they cannot do so in RW. And since rights to freedom and wellbeing are universal, the VWs' End-User Licensing Agreement (EULA) and Code must allow for actionable moral redress in rights violations within VW.

The general ethical principle therefore is that due to the OWS, avatars in all VWs, as modes of presentation or virtual extensions of their players in the RW have rights and corresponding obligations as both claimants and respondents in those VWs. Hence,

avatars as virtual representatives or modes of presentations of players have rights to freedom and wellbeing, prima facie as agents and absolute as persons, and those rights are actionable in VWs.

You will note that the above response to the objection raised in section 2.2 relies only and is facilitated by a very minimal account of the identity-relationship between avatar and player. In other words, the response to that objection does not rely on any controversial or problematic metaphysical account of the identity-relationship of avatar and player. Indeed, the *more of presentation* or *virtual extension* account I have used for the avatar and player relationship in my argument effectively bypasses any commitment to an independent or quasi-independent identity status for the avatar. But by being metaphysically silent or topic-neutral it is neither incompatible with such a position. I merely chose not to rely on any metaphysical identity theory for avatars in granting them rights, since none is required.

### *3.2. Conflict of rights*

In the event of a conflict of rights within a virtual world the conflict can in principle be resolved on the basis of Gewirth's subordinate argument for the Degree for the Necessity of Action Principle (DNA principle). The DNA specifies that in the event of a mutually exclusive conflict between agent A's generic rights (rights to freedom and wellbeing) and the generic rights of agent B, A's rights should take priority over those of agent B's rights when the objects of those rights, namely, freedom and wellbeing, are more necessary for purposive action for agent A than they are with regard to that of agent B [3].

### *3.3. Virtual role morality and universal public morality*

An important and overriding proviso is that the code and EULA of a virtual world must always be consistent with and not contravene the requirements of the Principle of Generic Consistency (PGC), especially as they apply to the absolute rights to freedom and wellbeing that all avatars/players have, by virtue of their dignity as persons. Thus a virtual world's Code or EULA is never justified in creating a *role morality* for itself that contravenes the universal morality supported by the PGC. The latter, because universal and foundational, being based on the PGC as the supreme principle of morality, will always take moral priority and override the *role morality* of any virtual world that allows for example, virtual rape or racial vilification, even though this might be allowed by the interration laws that are instantiated by the virtual world's Code and EULA. In other words, the interration laws of a virtual world must themselves be consistent with and not contravene the PGC.

### *3.4. Virtual rights are universal rights*

In conclusion, the rights of avatars in virtual worlds like the rights of their counterpart players in the real world are universal rights that apply always and everywhere. Thus an avatar and his counterpart player in the real world have the same universal rights to freedom and wellbeing irrespective of whether they reside in China, Kenya, Saudi Arabia, Iraq, Europe, America, Australia or anywhere else in the world or in cyberspace. Of course how those rights are used or applied to pursue individual and collective goals may vary from place to place from person to person. However, as

Gewirth's argument for the PGC and my reconstruction of it around the notion of dignity clearly demonstrates, since freedom and wellbeing are the necessary features of all action, they form the basis of universal rights to those necessary goods for all purposive agents, both real and virtual, for without them no purposive action would be possible, either in the universe or the metaverse.

## 4. Conclusion

The following is the conclusion of the paper in outline:

- Avatars have Rights to freedom and wellbeing, at least minimally, as Virtual Representatives or Modes of Presentation of their Players.
- Due to the Opacity of Worlds those rights are actionable in Virtual Worlds (VWs) since they cannot be actionable in the Real World (RW).
- Hence, RW and VWs are Morally Porous – there's no *Moral Magic Circle* [5] between RW and VWs.
- Virtual Rights (the rights of avatars) are Universal Rights and thus Global.
- Those rights are Prima-Facie with regard to purposive agency and Absolute with regard to personal dignity in both RW and VWs.
- Universal Public Morality (a universal morality based on the PGC) is applicable in both RW and VWs, and always overrides the Role Morality of any VW and those in RW when those two types of moralities come into conflict.
- Hence, the EULAs and Codes of Virtual Worlds must always adhere to UPM, particularly with regard to the absolute rights that avatars/players have to their dignity both personally and communally.

## Endnotes

[1] A full and detailed defence of the argument for the PGC against all the major objections raised against it by various philosophers can be found in Spence[4] (Chapters 1 to 3), Beyleveld 1991 and Gewirth [3].

[2] I use the terms *self-respect* and *dignity* interchangeably and with the intention that the terms be understood to apply both with regard to the individual person and to the community (ies) to which that individual person belongs. Thus a degradation say to one's race, nation or gender, would also be degradation to the individual persons that comprise that race, nation or gender, if it were perceived as such by those individual persons.

[3] A criminal, for example, sentenced to the death penalty has an absolute right to be executed with dignity. So although their prima facie rights to freedom and wellbeing with regard to their purposive agency are radically infringed qua agent they cannot be violated with regard to the preservation and maintenance of their dignity qua person.

[4] Peter Ludlow has conveyed to me in conversation his view that an avatar is no more than a different mode of presentation of the player, suggesting that an avatar does not in any sense hold a separate, albeit related, identity to that of their player

## References

[1] Richard A. Bartle. *Virtual Worldliness*, in Balkin M. Jack. And Noveck S. Beth (eds). The State of Play: Law, Games, and Virtual Worlds. New York: New York University Press, 2006. pp..31.
[2] Ralph Koster. *Declaring the Rights of Players*, in Balkin M. Jack. And Noveck S. Beth (eds). The State of Play: Law, Games, and Virtual Worlds. New York: New York University Press, 2006, pp. 55-56.
[3] Alan Gewirth. *Reason and Morality*. Chicago: University of Chicago Press, 1978, pp.48-128.
[4] Edward H. Spence *Ethics Within Reason: A Neo-Gewirthian Approach*. Lanham MD: Lexington Books (an imprint of Rowman and Littlefield), 2006, pp. 159-213.
[5] Edward Castronova. *The Right to Play*, in Balkin M. Jack. And Noveck S. Beth (eds). The State of Play: Law, Games, and Virtual Worlds. New York: New York University Press, 2006, pp. 79.
[6] Jack M. Balkin. *Law and Liberty in Virtual Worlds*, in Balkin M. Jack. And Noveck S. Beth (eds). The State of Play: Law, Games, and Virtual Worlds. New York: New York University Press, 2006, pp..91.
[7] Alan Gewirth. *Human Rights: Essays on Justification and Application*. Chicago: University of Chicago Press, 1982, pp..219.
[8] Edward H. Spence and Brett Van Heekeren. *Advertising Ethics*. Upper Saddle River, NJ: Pearson/Prentice Hall, 2005, pp. 1-16.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# On the Ecological/Representational Structure of Virtual Environments

Omar ROSAS*
*Department of Philosophy, University of Twente, The Netherlands*

**Abstract.** This paper introduces an alternative view of virtual environments based on an analysis of two opposing views: the Traditional View and the Ecological View. The Traditional View argues for a representational view of perception and action susceptible of being mapped onto virtual settings. The Ecological View, which is inspired by Gibson's ecological approach to perception, considers that perception and action are inseparable, embodied processes that do not imply mental representations. The alternative view put forward here claims for an articulation of the opposing views, namely the Ecological/Representational view of virtual environments, providing the notion and levels of representation implied in perceptual and agentic processes is functionally defined.

**Keywords.** Virtual environments, ecological approach to perception, Gibson, affordances, perception/action

## Introduction

Over the past 15 years, the growing development and multifarious applications of Virtual Reality Technologies have yielded a variety of desktop and immersive Virtual Environments (VEs), the implementation of which pervades different life domains ranging from everyday mobile communication, computer-based tasks, e-banking, collaborative learning, to specific uses in scientific research, medical care, and military industry among others. Such a rising development has brought about a bulk of literature coming from diverse disciplines and addressing different aspects of VEs. On the one hand, psychologists have studied people's perceptual, cognitive, affective, and behavioral responses to VEs and discussed the ecological validity of virtual reality as a research tool [1]. Sociologists and communication theorists have analyzed the impact of virtual communication and cyberculture on intersubjective practices and activities carried out in VEs [2]. Philosophers have examined the metaphysical and epistemological implications of VEs in terms of both their ontological status and their implications for people's perception and judgments of reality [3].

On the other hand, interdisciplinary research on VEs, mostly represented by the joint efforts of philosophers, psychologists, and cognitive scientists working on immersive environments, has focused on the phenomenon of *presence*, that is, the

* Correspondence address: Omar Rosas, P.O. Box 217, 75000 AE Enschede, The Netherlands. Tel.: +31 53 489-2308; E-mail: O.V.Rosas@gw.utwente.nl

feeling of "being there" inside the VE. Investigations into the nature of presence in VEs have raised issues about the adaptive dimension of perception and action in virtual worlds, and have been principally driven by two trends. The first trend, which may be called the Traditional or Rationalistic View, considers presence as a subjective, inner state (i.e., feeling) originated from an agent's perceptual immersion in and interaction with an external, digitally reproduced world (i.e., a virtual environment). It has been argued that the theoretical versions composing this trend still convey the old-fashioned subject/object dichotomy, and further the claim that for agents to successfully perceive and act in VEs they have to recruit *mental representations* in order to bridge the subjective and objective sides of the virtual experience [4].

The second trend, which is known as the Ecological View, draws principally on Gibson's [5] ecological approach to perception and regards presence as a unified phenomenon, stemming from the very interaction between an agent and its environment. On this view, presence is an agent's awareness of his existence in the (virtual) world, a kind of awareness that is claimed to arise from the intertwining of the agent's "virtual perception" and "virtual action". The agent is said to directly perceive the action opportunities or "affordances" provided by the VE, and to do so he does not need to recruit representations since the perceptual information required to act is already available in the very structure of the VE.

The debate opposing the defenders of each trend has mostly revolved around the explanatory advantages of one view over the other. Apart from some exceptions (see [6,7]), authors have paid little attention to the possibility of building bridges and working on the interfaces between their views. Yet on closer inspection, both views are complementary to each other to the extent that the ecological meaning of affordances can be soundly understood in representational terms. This certainly implies a re-examination of the concept and level of representation supposed to be implied in the perception of affordances as well as an elucidation of the reasons addressed by the defenders of the ecological view to rule out any representational features of perception.

In this spirit, my aim in this paper is twofold. First, I will argue that an ecological account of perception and action is not necessarily incompatible with a representational view of the mind. I will draw on three arguments coming from evolutionary studies and neuropsychological research to support this claim and defend an ecological/representational model of perception and action. Second, and by implementing the preceding claim, I will argue that VEs can be accurately accounted for within this model, and that such a model allows for a representational understanding of virtual and real affordances that is helpful for VE design.

The structure of the paper is as follows. The following section introduces and analyzes the basic source and assumptions of the Ecological View of VEs. The second section introduces the three arguments against an Ecological View that does not take into account representations in its explanation of perception and action. The third section capitalizes on and implements the outcomes of the preceding section to argue for an ecological/representational structure of VEs. The final section is devoted to some conclusions and issues for future research.

## 1. The Ecological View of VEs: Perception without Representations

Philosophical and psychological grounds for the Ecological View of VEs can be found in the papers by Flach & Holden [8], Zahorik & Jenison [4], Mantovani & Riva [9], Biocca [7], and Gross, Stanney, and Cohn [10] among others, issued in the journal *Presence*. Although these authors have been typically concerned with finding appropriate ways of elucidating the ontological and epistemological nature of presence, that is, whether it is a subjective feeling or an objective state facilitated by technological displays, this concern only represents the tip of the iceberg. What is deeply implied in their views is the idea that a clear understanding of how perception and action are effectively interrelated in real environments must drive research on the design and implementation of VEs, and that this understanding cannot be gained by relying on traditional conceptions of perception and action. In order to provide suitable grounds for this enterprise, these authors draw on a common theoretical source and share a common methodological assumption. The common source often invoked by defenders of the Ecological View comes from Gibson's ecological approach to perception. This approach is posited as a non-rationalistic, situated framework suitable to understand the intertwining of perception and action in real environments. The common assumption implies that VEs are isomorphic to real environments to the extent that the perception/action dynamics of the former can be accurately modeled on the ecological structure of the latter. Let us consider the implications of the source and the assumption in some detail.

### *1.1. Gibson's ecological view*

According to Gibson's ecological approach to perception, perceiving is the direct process of picking up information from an already informationally rich environment. The nature of the information provided by the environment is not to be confused with proximal stimulation in the form of sense data, but rather to be viewed as action-guiding properties or dispositions of the objects inhabiting that environment. Such properties are what Gibson calls *affordances*, that is, specific invariants the perception of which is meant to support an organism's action. On Gibson's view, affordances are properties integrating both agent and environment into an embodied system, and the nature of this embodied relation entails the claim that to perceive the world is to co-perceive oneself.

Gibson's conception of perception as implying an intimate relation between exteroception and proprioception is meant to get rid of dualistic views in which the agent is thought to perceive or see the world through mentally processing raw sense data provided by the organs of its body. As he used to argue, agents do not see their environment with the eyes but with an embodied system composed of the "eyes-in-the-head-on-the-body-on-the-ground". This embodied system is conceived by Gibson as a functional unit that cannot be reduced to a juxtaposition of discrete anatomical parts. For Gibson, the idea that perception is the work of functional perceptual systems rules out any conception aimed at dividing perceptual experiences into subjective and objective dimensions. On his view, it is clear that the ecological complementarity between agent and world is not separable.

Furthermore, Gibson draws a clear distinction between perceptual senses and perceptual systems, attributing to the latter complex and functional operations that go beyond the mere registration of stimuli. A sense has just receptors whereas a system has organs that orient, explore, adjust, and come to equilibrium at a given level of subordinate or superordinate functioning. Perceptual senses are conceived as somewhat rigid mechanisms, grounded on a repertoire of innate sensations while the achievements of a perceptual system are susceptible to maturation and improving via learning.

Gibson's reluctance to ground perception on representations originates from his firm conviction that the pick up of relevant information is not filtered by an agent's mental models and processes, and that perceiving is an achievement of the whole individual, not an appearance in the theater of his consciousness. Perceiving in this sense concerns "keeping-in-touch" with the environment and provides the notions of situatedness, embodiment, and embeddedness with a full ecological meaning.

By assuming a radical position against mentalistic explanations of perception, Gibson argues that the term "representation" is misleading. For him, there is no literal re-presentation (as in a photograph) of an earlier optic array perceived in the environment; only some of its invariants can be preserved, but that is all. His dissatisfaction with philosophical and psychological approaches to representation targeted conceptualizations of representation as "pictures in the head", which are supposed to tie—in a rather obscure way—the objective (sensory) and subjective (mental) contents of experience. A major flaw in this view has been often addressed, namely that it leads to metaphysical perplexities like a surreptitious "inner eye" or "homunculus" whose function is to verify the consistency between the object and its corresponding 'picture', and guarantee the appropriate matching between objective and subjective experiential contents. However, beyond his rejection of pictorialist accounts of representation, Gibson's negative attitude towards representational accounts of perception and action in general is but a symptom of his deep skepticism about the explanatory promises of cognitive and computational models of the mind championed by several of his contemporary colleagues.

### *1.2. The isomorphism between real environments and VEs*

As stated above, the common assumption of the Ecological View is that VEs are isomorphic to real environments to the extent that the perception/action dynamics of the former can be modeled on the ecological structure of the latter. This assumption implies that (a) "virtual perception" and "virtual action" are inherently related via "virtual affordances", and (b) agents navigating through VEs *directly* perceive the action opportunities furnished by virtual objects and virtual agents. It should be noted here that, though the isomorphism between real environments and VEs is *mutatis mutandis* structurally and functionally valid, it is nonetheless partial. This is so because, to date, some real-world affordances cannot be mapped onto their virtual counterparts. For example, a virtual fruit affords, say, "grasp-ability", manipulability, "throw-ability", but not edibility. Equally, a virtual glass of water affords "reach-ability" and "break-ability", but not "drink-ability". However, these extreme cases of technological irreproducibility of affordances do not undermine the basic assumption of the Ecological View since the essentials of perception and action in real environments can be reliably applied to VEs.

This reliability or ecological validity of VEs is largely contingent on designers' accurate understanding of the nature of affordances. For a VE to be meaningful in the ecological sense, programmers and software developers have to provide users with sensible relations between actions, affects and effects, no matter whether the VE is designed for psychological therapies, collaborative learning, or even video games in which agents can fly, resuscitate, clone themselves, or metamorphose into other creatures. Such relations represent a design commitment to provide a satisfacing compromise between virtual events, available actions, and users' expectations. Furthermore, designers of VEs must be aware that the meaning of virtual actions is not just a construct of the user's mind. Sufficient and clearly detectable physical, semantic, and cultural information has to be provided by the very structure of the objects and agents inhabiting a given VE in order to afford users the possibility of choosing between alternate patterns of action marked by explicit degrees of freedom.

Defenders of the Ecological View of VEs consider Gibson's ecological approach to perception as a compelling framework for disposing of representations and validating a direct, embodied account of perception and action in VEs. Animated by this ecological impetus, they claim that the Traditional View is inadequate because of its artificial separation between objective (sensory information) and subjective (cognitive processing) dimensions of perceptual experiences, thus conveying the idea of an agent interacting in a distal, disincarnated way with his (real and/or virtual) surroundings.

Nevertheless, it is necessary to consider that in view of recent functional accounts of perception and action originated from evolutionary, psychological, and neurophysiological perspectives, Gibson's ecological approach needs some critical scrutiny in order to test its basic tenets and concepts for theoretical accuracy and explanatory power. The outcomes of such a scrutiny will certainly have an impact on the way the Ecological View conceives of the structure of VEs. In the following section, three arguments coming from the aforementioned perspectives will be analyzed.

## 2. Three Arguments for the Need of Representations

Recent evolutionary, psychological, and neurophysiological research on cognition, has delivered interesting models for understanding perception and action in accurate representational ways. For the most part, these models aim at articulating preceding and, to some extent, competing views of perception/action into a theoretical framework that allows for cognitive and embodied explanations of the interactions between agents and their environment. Here, we will briefly examine three arguments in favor of mental representations advanced by these models.

### *2.1. The environmental complexity thesis*

By adopting an adaptationist and pragmatic view on the evolution of cognition, Godfrey-Smith has argued that the complexity of the environment drives the need for agents to develop complex cognitive resources and capabilities. His basic claim is condensed in the *Environmental Complexity Thesis*: 'The function of cognition is to enable the agent to deal with environmental complexity' [11]. This thesis implies two causally interrelated corollaries concerning the relations between agent and

environment: a) the more complex the environment, the more elaborate is the behavioral repertoire required to deal with the environment, and b) the more elaborate the behavioral repertoire, the more complex are the cognitive capabilities and mental representations needed.

Godfrey-Smith understands environmental complexity in terms of 'heterogeneity'. This means that environments can be varied, diverse, changing, and offering the agent opportunities to face a lot of different states. Yet the notion of heterogeneity underlying environmental complexity is not an all-or-nothing condition: any environment can be heterogeneous in some respects, and homogeneous in others. The complexity of a given environment can be assessed in different states at different times, rather than the same state all the time. Different states represent changes in the environment, changes that largely amount to what Gibson refers to as *environmental events*: modifications or transformations of the environment's objects, surfaces, and layout that impose constraints on the organism's perceptual and agentic activities. Furthermore, Godfrey-Smith considers that the complexity properties of environments are to be regarded as objective, agent-independent properties of which only a few will be relevant to any given agent. Whether a specific kind of complexity matters or is relevant to an agent will depend on what the agent is like: on its physiology, cognitive endowment, needs, and habits.

Furthermore, the objective status of complexity properties as well as their agent-relative relevance implies that for agents to track and deal with complexity properties, they also need to develop a kind of complexity, namely *cognitive flexibility*. Cognitive flexibility is seen as an agent's adaptive response to challenging conditions of their environments, a kind of adaptation produced by evolution to enable agents to perceive what environmental invariants persist, and what have either changed or gone out of existence, in order to regulate their behaviors and seize environmental 'opportunities' in successful ways. Cognitive flexibility also implies that agents possess and develop complex representational architectures that can be constantly improved given the agent's potential to learn. This latter point shares with Gibson's notion of a perceptual system the idea that perceiving is a process open to improvement given the susceptibility of perceptual systems to maturation and learning.

### *2.2. The evolution of decoupled representations*

In an adaptationist spirit similar to that of Godfrey-Smith's, Sterelny has provided an evolutionary explanation of the consequences of different informational environments on the evolution of cognitive systems [12]. His baseline for developing his claims is that organisms have mechanisms that mediate specific adaptive responses to features of their environment by registering specific environmental signals telling them of the presence of those features. Yet although the signals of informationally rich features of the environment can be accurately perceived by an organism, this may not be always the case.

Sterelny points out that agents living in complex and changing environments can develop robust tracking systems. Unlike simple detection systems, robust tracking emerges as a flexible adaptive response to the informational structure of a given environment. When considered from an informational perspective, environments can be of three kinds: *transparent*, *translucent*, or *opaque*. Transparent environments are so stable that they allow for adaptive responses by using reliable specific cues. Yet if the ecologically relevant features of the environment map in complex ways onto the cues

an agent can detect, this agent is living in a translucent environment. In order to cope adaptively with a translucent environment, an agent has to develop functionally flexible ways of perceiving salient features enabling it to discriminate a functional category (e.g., food, shelter, mates, etc.) via multiple perceptual channels (e.g., vision, smell, auditions, etc.). Finally, if the environment becomes so unstable that even functional flexibility in tracking salient features is bound to misfire, then the agent is living in an informationally opaque environment. To be sure, environments are typically heterogeneous in that they are transparent with respect to some features, translucent with respect to others and even opaque with regard to still others. In a clearly ecological spirit, Sterelny claims that the epistemic character of an environment is the result of an organisms adapting to its physical circumstances, tuning its perceptual channels to pick up information that the world provides for free.

To supply an explanation of how agents adapt and evolve in complex environments, Sterelny has postulated the evolution of *decoupled representations*. According to him, decoupled representations are "internal cognitive states which (a) function to track features of the environment, and (b) are not tightly coupled functionally to specific types of response" [12]. Decoupled representations are thus conceived as 'fuel for success' since they constitute a flexible information database that enables the agent to carry out perceptual and agentic activities without being tied to specific behaviors. Moreover, decoupled representations evolve along with response breadth, which means that decoupling is a matter of degree, developing from an increasing flexibility in the use of information agents pick up.

### *2.3. The theory of event coding*

The basic claim of the Theory of Event Coding (TEC) advanced by Hommel et al. [13], is that perception, attention, intention, and action share, or operate on, a common representational domain. This theory draws on both cognitive and ecological views of perception and action, and attempts to articulate theoretically and empirically the central tenets of each view.

Unlike most theories that assume a functional dichotomy between perceptual codes and action codes, the TEC denies that perceiving a stimulus and planning a voluntary action are distinct processes operating on completely different codes. For Hommel et al., perception and action are equivalent insofar as they are alternative ways of internally representing interactions between ecological events and the perceiver/agent. Several claims can be distilled from TEC. First, perceiving is seen as a process of actively acquiring information about the perceiver/environment relationship, a process that implies allocating cognitive resources (e.g., attention, memory) to salient features of the environment. Second, the process of perceiving presupposes and affords active behavior, and action in turn affords perceptual information. Third, to the extent that environmental actions can be considered as coming into being by anticipating their distal effects, perceived events and their consequent affordances are coded and stored together in one common representational domain. Fourth, the functional equivalence between perceptual and action codes stems from the fact that both kinds of code refer to external events: the codes representing the intended action features are already activated in the course of perceiving the stimulus, so underlying the perception/action dynamic relation.

Furthermore, TEC posits that feature codes—that is, the codes that represent the distal features of an event—are not specific to a particular stimulus or response, but

register information from various sensory systems and modulate various motor systems. In a line similar to the evolution of decoupled representations put forward by Sterelny, TEC considers that salient features of the environment can be perceived through more than one sensory modality and it is the common coding of perception/action that integrates this information. Interestingly, TEC also claims that the dimensions feature codes refer to need not always be properties like color or shape, but can also be as complex as 'edibility', 'graspability' or 'sit-on-ableness" in a Gibsonian sense.

Although essentially conceived as a cognitive model for perception and action applicable to behavioral data, TEC also find support in recent neuroanatomical and neurophysiological evidence for brain modules shared by perception and action planning. For instance, the discovery of mirror neurons has provided support for the functional overlapping of perceptual and action-related codes. Mirror neurons discharge during the performance of goal-directed actions and the perception of actions performed by others [14]. These neurons have been identified in the ventral premotor and posterior parietal cortices of monkeys, and a number of functional neuroimaging studies with humans documented the selective recruitment of homologous cortical regions that implement perception/action representations in human premotor and parietal cortices.

Taken together, these three arguments speak in favor of a representational view of perception and action (and, by extension, of the mind) which is overtly compatible with the central tenets of Gibson's ecological view. The next section will be devoted both to an articulation of the representational with the ecological view of perception/action and to a targeting of relevant implications of this articulation for the Ecological View of VEs.

## 3. Redefining the Ecological/Representational Structure of VEs

Recall that the core assumption of the Ecological View of VEs is intended to support the claim that agents can engage in purposeful perceptions and interactions inside VEs in as much the same way as they do in real environments. For the defenders of this view, Gibson's ecological conception of perception has proven to be a suitable theoretical means both to map real perception and action onto virtual experiences and to explain the advantages of this view for the design and implementation of VEs. On this view, agents are said to navigate through VEs by engaging their perceptual systems and agentic capacities in a direct way, without being bound to draw on mental representations of virtual objects and/or virtual agents.

Yet as we have seen in the preceding section, the story about the ecology of perception/action without representation is far from being uncontroversial. The point is not that the Ecological View of VEs is fundamentally flawed or false; rather, it is that its common source and core assumption are in need of some conceptual and methodological fine-tuning. The arguments introduced above provide us with cogent reasons to carry out such a fine-tuning, and suggest that an ecological view of perception and action is compatible with a representational view of the mind, provided the concept of representation is understood in a functional, non-'pictures-in-the-head' sense.

Let us begin by noting that representations can be considered to enter the ecological view of perception/action at a functional level as the basis of an agent's

cognitive schema of individuation. Such a schema corresponds to a set of evolved representational capacities, contingent on the agent's cognitive endowment and allowing it to adaptively perceive and deal with its environment in terms of meaningful arrays of perceptual invariants. The adaptive perception of these arrays implies that affordances have a representational dimension anchored in the intrinsically normative character of action opportunities (physical regularities of the environment or social rules of a group). Moreover, this schema of individuation is not conceived as a purely conceptual template to be stamped on the world, but rather as an active mechanism assembling perceptual and action-related codes to enable the agent to be aware of both his situation in the environment and the opportunities furnished by objects and other agents. Far from being a mere reductionist strategy, the fact that cognitive and neuroanatomical evidence lends support to the emergence and operations of this schema of individuation gives us reliable arguments to defend an ecological/representational view of perception and action, in which both the whole acting subject and his embodied brain activity play crucial roles in picking up of invariants and having meaningful perceptual experiences.

Now, as far as the Ecological View of VEs is concerned, this fine-tuning of its common source has two basic implications for the common assumption. First, to the extent that humans have not naturally evolved in technologically reproduced virtual settings, it is reasonable to expect that they capitalize on their natural representational abilities to perceive and act in VEs. Second, apart from being representational at a functional level (given that they are experienced by recruiting cognitive capacities), VEs are also representational at an epistemic level, namely as deliverers of knowledge representations.

VEs provide agents with three kinds of knowledge representations: analogical, propositional, and procedural [15]. Analogical representations preserve properties of objects and events in an intrinsic manner and keep their same inherent constraints. In this sense, an icon on the computer screen, a labyrinth in a video game, a virtual fruit, or even an avatar, all are analogical representations. They are ontological reproductions or simulations of real objects and events, and are designed to functionally keep the latter's physical attributes and action opportunities. Propositional representations preserve the structure of objects and events extrinsically. When navigating through a given VE, linguistic indications such as "Enter Here", "Members Only", "User Name", or "Password" provide users with information relevant to carry out specific patterns of action. Here objects and events are extrinsically, linguistically coded to afford clear perceptual and active engagement. Finally, procedural representations provide users with specific rules to accomplish particular actions within the VE. These representations can take the form of a set of instructions to sign up for a given website, to master an avatar's movements, or to find a way out of a virtual maze.

It is clear that an appropriate understanding of these representational issues is crucial for a meaningful design and implementation of VEs. For the way the functional and epistemic levels of representation are brought inside a given VE determines its entire epistemic structure. The essential difference between real and virtual settings rests on a platitude worth recalling: VEs are artificially reproduced worlds whose entire affordance-related and epistemic structure is largely a matter of designers' intentions and goals. In this sense, it is up to VE developers to draw on an ecological/representational understanding of perception and action in order to realize that their design of virtual worlds can be as transparent, translucent or opaque as their real counterparts.

## 4. Conclusions

We have seen that, despite the claims of the Ecological View of VEs, a representational account of perception and action, either in real or virtual environments, is compatible with an embodied, ecological conception such as that championed by Gibson. This compatibility resides in the fact that, *pace* Gibson, not any representational account of perception necessarily implies a pictorial conception of representation or a dualistic view of perceptual experiences. Arguments from evolutionary, cognitive, and neuroanatomical studies have proven to be helpful to elucidate a functional, representational model of perception and action, and to favor an ecological/representational explanation of the structure of VEs, without having to betray the ecological, and certainly theoretically valuable, view of embodied perception. Accounts of cognition and environmental complexity, decoupled representations, and common coding for perception and action, have yielded compelling reasons for arguing that perceptual and agentic processes can be ecological as well as representational.

The implications of this for the design of VEs have been explored. The ecological/representational analysis of VEs provided here, makes clear that for meaningful use of VEs, users draw on their natural cognitive, representational endowment to accurately track structural and functional correspondences between real and virtual worlds. This point has been argued to be useful for VE designers, since it is their understanding of how affordances and significant features are perceived and engaged in that will make a VE epistemically transparent, translucent or opaque.

## References

[1] Y. A.W. de Kort, W. A. IJsselsteijn, J. Kooijman, and Y. Schuurmans: Virtual Laboratories: Comparability of Real and Virtual Environments for Environmental Psychology. *Presence* (2003), 12 (4): 360-373.

[2] D. Tofts, A. Jonson, and A. Cavallaro: *Prefiguring Cyberculture: An Intellectual History.* (2003), Cambridge University Press.

[3] A. Borgmann: *Holding onto Reality: The Nature of Information at the Turn of the Millennium.* (1999), The University of Chicago Press.

[4] P. Zahorik and R. L. Jenison: Presence as Being-in-the-World. *Presence* (1998), 7(1): 78-89.

[5] James J. Gibson. *The Ecological Approach to Visual Perception.* (1986) Lawrence Erlbaum, Hillsdale, NJ.

[6] T. B. Sheridan: Descartes, Heidegger, Gibson, and God: Toward and Eclectic Ontology of Presence; *Presence* (1999), 8(5): 551-559.

[7] F. Biocca: Inserting the Presence of Mind into a Philosophy of Presence: A Response to Sheridan and Mantovani and Riva. *Presence* (2001), 10(5): 546-556.

[8] J.M. Flach and J.G. Holden: The Reality of Experience: Gibson's Way. *Presence* (1998), 7(1): 90-95.

[9] G. Mantovani and G. Riva: Building a Bridge between Different Scientific Communities: On Sheridan's Eclectic Ontology of Presence. *Presence* (2001), 10(5): 537-543.

[10] D.C. Gross, K.M. Stanney and J. Cohn: Evoking Affordances in Virtual Environments via Sensory-Stimuli Substitution. *Presence* (2005), 14(4): 482-491.

[11] P. Godfrey-Smith: *Complexity and the Function of Mind in Nature.* (1996), Cambridge University Press, Cambridge.

[12] K. Sterelny: *Thought in a Hostile World. The Evolution of Human Cognition.* (2003), Blackwell Publishing, Oxford.

[13] B. Hommel, J. Müsseler, G. Aschersleben, and W. Prinz: The Theory of Event Coding (TEC): A Framework for Perception and Action. *Behavioral Brain Sciences* (1998), 24: 849-937.

[14] C. Lamm, M.H. Fischer, and J. Decety: Predicting the Actions of Others Taps into One's Own Somatosensory Representations. A Functional MRI Study. *Neuropsychologia* (2007), 45: 2480-2491.

[15] T.P. McNamara: Knowledge Representation. In Robert J. Sternberg (Ed.) *Thinking and Problem Solving*, Academic Press, London, 1994: 81-117.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*


# The Dynamic Representation of Reality and of Ourself between Real and Virtual Worlds

Lukasz PIWEK
*Department of Philosophy, University of Glasgow, UK*

**Abstract.** There seems to be a difference in the way we interact with reality and a reality experienced while playing computer games. I will argue that one of the most important features that distinguishes external world (or Open Reality) from reality experienced while playing computer games (or Closed Reality) is the degree of complexity, that is, the richness of the stimuli and the number of options available. One of the main consequences of the lower complexity of Closed Reality is that playing computer games triggers different cognitive alterations in an effortless and automatic manner. The question I ask is what really changes in our cognitive processing when we play computer games. One of the answers is that there is a change in the agent's cognitive representation of reality. Additionally I will suggest that there seems to be a change in the cognitive self while playing avatar-based computer games. I will discuss the last point in the brief context of identity problem and possible psychological implications.

**Keywords.** Computer games, Virtual Reality, cognitive representation, self, avatar

## Introduction

This article will mainly focus on the problem of cognitive changes in a person who plays computer games. While ‘agent’ refers to a being that a person typically is in his or her everyday life (a person representation in external reality), ‘avatar’ refers to a being that a person is in a computer game world (a person representation in virtual reality). While there is always only one agent, this agent can have many different avatars.

I will discuss the theory of an Open and a Closed Reality together with notion of an open and a closed cognitive representation. I will also refer to psychological findings related with changes in agent’s cognition while playing computer games. In the final part, I will describe a theory of Cognitive Self in the context of avatar-based computer games.

I define computer games as all video games, mobile games and any other forms of games based on Active Virtual Reality (Active VR). Active VR is a defined virtual environment where an agent has distinctive goals, missions, and aims (e. g. *World of*

*Warcraft*, *The Sims*). Passive Virtual Reality (Passive VR) is a defined virtual environment where the agent does not have distinctive goals, missions, and aims (e. g. *Google Earth* or internet browsing in general). I will only focus on Active VR in this article.

## 1. Virtual Reality and Interface

### *1.1. "The map that precedes the territory"*[1]

> "Today abstraction is no longer that of the map, the double, the mirror or the concept. Simulation is no longer that of a territory, a referential being, or a substance. It is the generation by models of a real without origin or reality: a hyperreal. The territory no longer precedes the map, nor does it survive it. It is nevertheless the map that precedes the territory – precession of simulacra – that engenders the territory, and if one must return to the fable, today it is the territory whose shreds slowly rot across the extent of the map. It is the real, and not the map, whose vestiges persist here and there in the deserts of the real itself (…)" [1].

In this paragraph from *Simulacra and Simulation* Jean Baudrillard refers to Borges' metaphor of the cartographers of the Empire who draw up a map so detailed that it ends up covering the whole territory [1]. Nevertheless, his simulacrum is gaining a new meaning in our times. The impact of computer games is growing beyond the line between virtual world of entertainment crossing into the real world of economy, culture, politics and education. In games such as *Second Life*, *Entropia Universe* or *World of Warcraft* agents can buy and sell virtual land for real money, build their own islands, design buildings and other elements of reality or buy and trade items from the game on internet auctions for real money. Universities are organizing lectures and conferences, companies running charity events and concerts played by real DJs – everything in *Second Life* [2]. An outbreak of a deadly disease ("corrupted blood") in *World of Warcraft* was analyzed by scientists from Tufts University as a possible insight into real life epidemics and human behavior under such conditions [3]. Statistics in the US show that gamers devote more than triple the amount of time playing games each week than exercising or playing sports, volunteering in the community, religious activities, creative endeavours, cultural activities, and reading [4].

Computer game developers aim to provide increasingly entertaining experiences by multiplying interfaces and increasing complexity and viability of games (e. g. new generation consoles graphic processors, 3D Dolby Surround sound systems or *Nintendo Wii* motion controller). Simultaneously they try to find a way to boost interactivity and attach a game to a deeper level of agent experience. *Second Life* economy, *The Sims* reality where an agent designs the whole existence of virtual families or *World of Warcraft* with expanding character-building paths and hierarchy provide only a few examples. The internet seems to be a complimentary technology that helps to achieve a higher social interaction and therefore increases interactivity. Potentially, the ultimate aim would be to create a virtual reality in which an agent could have a full sensory experience. Such reality would be parallel to ours in a social, cultural, political and economical sense. Such creation would be similar to those described by William Gibson or Neal Stephenson in their novels[1]. In such context, there would be no

[1] That is William Gibson *Neuromancer* (1984) and Neal Stephenson *Snow Crash* (1992).

distinction between an external world and a virtual world. However, at the current stage of development of human technology there is still a long way to achieve such an aim.

### *1.2. Direct and indirect interface*

What is the difference between the cognitive experience of this world and the virtual world for an agent? Is cognitive representation of this world different from a representation of a virtual world? In some sense, it is the same. An agent experiences emotions from movies as he does experience emotions triggered by computer games. However, even considering above, a game is still a simulation (even if it "covers the whole territory" [1]). An agent (at least the majority of them) can discriminate between playing tennis with *Nintendo Wii* in virtual world court (even using motion controller) and playing tennis in an external world court. There are simple perceptual differences, sensations that vary between those two realities. There is still much more sensory experience available in external world than in virtual world. It is mainly due to the fact that an agent still experiences virtual reality via indirect interface and external world via direct interface.

When an agent is playing computer games, even using the most sophisticated devices, he is still limited to indirect interface. Indirect interface – that is a keyboard, a mouse, a pad, a joystick and a computer screen, a mobile or video screen – which allows an agent to interact with the computer game environment indirectly. Even systems available non-commercially like Virtual Reality (VR) goggles, are still unreliable and limited compared to biological vision. So an agent will have all sensual experiences with virtual reality via an indirect interface – he is not directly acquainted with it. In contrast, in the external world it is accessible more or less via a direct interface. The only filter that an agent has in experiencing an external world is his own brain and nervous system. In experiencing a virtual world, there is another filter, which is an interface device.

A direct/indirect interface should not be confused with the direct/indirect realism. A direct/indirect interface only refers to a problem of difference in access to an external/a virtual reality while the direct/indirect realism refers to global perception of the reality. An agent cannot experience virtual objects directly via senses (direct interface) - at least at this stage of technological development. Therefore, when an agent is driving a BMW car in *Second Life* virtual Florida, even using extraordinary vision and sound system with an artificial wheel that can conduct vibrations, he will still experience his BMW via an indirect interface. It will not provide the same sensory experience as driving a real BMW across the real Florida coast.

## 2. Dynamic Changes in the Representation of Reality

### *2.1. Degree of complexity and the rule of simplicity*

An agent can act in two realities – an open and a closed reality. An Open Reality (OR) is an external world. A Closed Reality (CR) is a virtual world experienced by an agent while playing computer games. What distinguishes OR from CR is a degree of complexity. **Degree of complexity** is defined as the richness of the stimuli and the number of actions-options available. The degree of complexity for a Closed Reality seems to be lower than for an Open Reality because (i) a Closed Reality provides less

potential sensory stimulation overall than an Open Reality, and (ii) a Closed Reality is more limited in the number of action-options. An agent has access to a Closed Reality only via indirect interface. What follows, a Closed Reality provides less complex environment and less potential stimulation than an Open Reality. Therefore, the most basic cognitive implication is that a Closed Reality is much easier to organize perceptually – it is simpler than an Open Reality.

> "The cognitive apparatus finds patterns in the data that it receives. Perception involves finding patterns in the external world, from sensory output" [5].

There are several lines of evidence in cognitive neuropsychology that support the claim that humans tend to prefer simplicity and they can typically process it quicker than more complex system [5]. Items with simpler description are typically easier to detect in noise [6]. In addition, the vast range of phenomena in perceptual organization, including Gestalt laws of closure, also supports simplicity theory [7]. An agent has an automatic tendency to complete and finish tasks in any defined system of reality [8] and has a preference for less complex systems. A Closed Reality is such a system – a reality that itself is a task to close, and has predefined goals, objectives and patterns of activity.

Structures in a Closed Reality are designed to be interconnected and as simple as possible to boost the learning curve for an agent interacting with it. An agent seeks an environment that he can control with a mouse-click and few keyboard buttons. The existence of an agent in Closed Reality is reduced to 'shortcuts' – it is limited, simplified and well structuralized (even in the sense of easy access, iconic graphics and easily remembered elements of interface). Limited interface brings comfort and easiness in mastering it and after a short time it becomes intuitive and effortlessly automatic. Finally, the simplicity of a code for a stimulus quantifies the amount of structure uncovered in that stimulus [5]. The more structure people can find in a stimulus, the easier they find it to process and remember and the less random it appears to them [9].

### *2.2. Open and closed cognitive representation*

An agent experiences different cognitive alterations to deal with constant changes in reality on a daily basis. For example, there are typically some general mood and patterns of thoughts related with different semantic categories, various states of awareness and attention. Therefore, an agent has different cognitive parameters in different moments of his everyday functioning. His cognition is framed within the cognitive representation of reality he is experiencing at any given moment.

Cognitive representation can be defined as a sensory-perceptual global description of a reality that adjusts the cognitive system to the complexity of a given reality. Cognitive representation is a dynamic, global, mental process that is functionally related to sensory organs and subjective perception. There are two possible states of cognitive representation, open and closed, open in an Open Reality and closed in a Closed Reality. When an agent connects to a Closed Reality his cognitive representation swaps from an opened to a closed reality, which causes a number of global changes in his perception and cognition.

### *2.3. Cognitive alterations in a Closed Reality*

What are the cognitive changes that an agent experiences when entering a Closed Reality? Typically, an agent who starts playing any online game is confused with the user interface, profile descriptions, shortcuts and functionality of the different options. Usually, an agent arrives at a training area, to make his adaptation smooth. It is similar to quarantine, where every agent learns the basics of the game. Typically, it takes him/her just a few hours to get used to essential parameters, and to understand the mechanics of the game. After that, he is just mastering his avatar and finding his own individual way of moving around the virtual space. It all occurs very quickly, an agent starts creating a detailed representation of all locations, parameters and categories. After a few weeks of playing, an agent knows exactly where to go to get sword X to kill the monster in location Y and get a reward in location Z.

The initial stage is based on mastering motor coordination and habituation to user interface in the game. The easier and more intuitive the interface is, the less time it seems to take to master the interface. However, those claims are only based on observation and more empirical study needs to be done to establish what the patterns are in interface habituation and learning.

Chou and Ting observed that an agent playing a computer game typically (i) is strongly focused, (ii) has a clear perspective of completing well-defined goals, (iii) receives immediate feedback on attempts to cope with obstacles, (iv) has a strong sense of control, and (v) experiences an altered duration of time [10]. Such experiences are specific for the cognitive state of flow [10]. Csikszentmihalyi defines the flow as:

> "mental state of operation in which person is fully immersed in what he or she is doing, characterized by a feeling of energized focus, full involvement, and success in the process of the activity" [11].

In his earlier definition, Csikszentmihalyi sees flow as a "shift in a mode of experience" when people "become absorbed in their activity" [12]. As Hoffman points out in his study on Virtual Reality analgesia:

> "human attention has been linked to spotlight, allowing us to select some information to process and to ignore everything else, because there is a limit to how many sources of information we can handle at one time" [13].

Hoffman suggests that spotlight attention increases with increased interactivity and richness of the virtual environment. He showed how patients with severe body burns could distract themselves more efficiently from pain when playing computer games, and that such alleviation improves with increased complexity of the interface and environment [13]. He used a virtual environment *SnowWorld* (in which agent experiences illusion of flying through an icy canyon with a frigid river and waterfall, as snowflakes drift down, and they can shoot snowballs at snowman, igloos, robots and penguins) to boost the effect of distraction from burns pain [13].

The effects Hoffman described are consistent with the findings of Chou and Ting on the state of flow [10]. Hoffman suggests that spotlight attention is a major factor that causes a following effect of distraction from pain and he points out that the effect was stronger for patients using stereoscopic goggles compared with those playing computer games on screen [13]. In addition, attention is certainly an important factor,

in Csikszentmihalyi notion of flow [12], but I would argue that Closed Reality experienced on flat screen could potentially be as immersing as a Closed Reality experienced via stereoscopic goggles. In *SnowWorld* an agent does not have any distinctive avatar – an agent is just 'himself' flying around the frozen world. It could potentially make a difference if the degree of complexity of SnowWorld would be increased by such features as the presence of an avatar, with skill development, fulfilling the quest, sophisticated AI interactions, exploring new areas and multiplayer option [14]. Agents usually report such factors as important in improving their gaming experience, beside high-quality realistic graphics and sound [14]. Moreover, when the complexity of a Closed Reality is expanded by the presence of distinctive avatar, this could cause a temporary alteration in an agent's Cognitive Self.

It is important to note, an increase in complexity occurs within a frame of Closed Reality that is limited by a richness of the stimuli and a number of action-options available in comparison with an Open Reality. Therefore, an increase in complexity within a Closed Reality does not contradict the rules of simplicity. Simplicity still applies because a higher degree of complexity in a Closed Reality is still not large enough to 'open' this reality in the perceptual and cognitive sense.

### *2.4. Relation between an agent and an avatar*

There are very basic and simple games, like *Pacman* or *Space Invader* that can certainly cause some temporary changes in an agent's cognition. Such games are extremely basic Closed Realities, with only a small number of action-options available. In modern online games such as *Second Life* or *World of Warcraft* the complexity is much higher, therefore there are probably more ways those games are impacting an agent's cognition. One of the elements that seem to be crucial from both psychological and philosophical perspective is the presence of an avatar in the game.

Waelbers and Spence claim that an agent's avatar is an "echo of their true selves" [15]. They also stress that it is an improved reflection of an agent's real identity. Certainly an avatar is a representation; an agent can clearly visualize and influence different attributes, skills, randomize items or powers. However, it is doubtful that there would be a straightforward similarity between an agent and his avatar. An avatar is not simply a simulation or even a simulacrum. If an agent would create seven avatars in five various games they would probably be very different from each other. That would simply be a consequence of different game mechanics (that an agent has different options for avatar creation in *Second Life* than in *World of Warcraft*). Further, there are psychological issues related with the way an avatar reflects an agent, and such variables as level of neuroticism or personality type would probably predict what kind of avatars an agent is creating[2]. Finally, different agents have different reasons for entering a Closed Reality, varying from leisure and distraction to more sophisticated social, cultural or economic purposes. Nevertheless, an agent who creates an avatar is still interacting with it. Avatar represents agent's actions in a Closed Reality where he experiences different cognitive changes. The suggestion is that the interaction between an agent and avatar has an impact on dynamic representation of an agent's Cognitive Self.

[2] However, this aspect is still hypothetical and more psychological research needs to be done to evaluate such claims.

### *2.5. Cognitive Self*

The self is here defined as a "mental thing, ontically distinctive that is a subject of experience and has a certain character" [16]. A Cognitive Self is a cognitive representation of the self that an agent is linked up with at any given moment.

When an agent plugs into Closed Reality and he adjusts to a closed cognitive representation, then we can assume that there will also be a cognitively closed representation of self. Such a difference could explain why an agent is experiencing the cognitive alterations immediately and non-inferentially while playing computer games (for example, changes in attention, motivation, goals and strategies, sense of control [10]). For example, writing an article requires a number of different activities in the way to achieve a finished copy and each of those activities requires different cognitive processes. Certainly, an agent can experience a state of flow while writing. However, there is a significant number of simultaneously occurring processes in an Open Reality, and that can simply cause distraction. That is how 'cognitive delays' like procrastination and attention distraction occurs, and these prevent the state of flow from occurring immediately. In contrast, it is hard to imagine an agent who 'procrastinates' while playing computer games or is distracted during this activity. It is paradoxically because Closed Reality is an ultimate distraction itself (as Hoffman suggested [13]), and therefore an agent's closed cognitive representation can not just 'open' toward any 'procrastination' within an activity in Closed Reality. When playing computer games, it seems to be ultimately easy to lose oneself in this activity. Such a feature is also the key to a flow-like experience [11].

Therefore, the change between Closed and Open Cognitive Representation is a dynamic process that also affects a Cognitive Self. This has been shown in many psychological experiments and theories. Examples include Higgins' self-discrepancy problem [17], Baumeister's self-escape theory [18] and Marcus' theory of possible selves [19], which shows that our self-representation is very liable and vulnerable to frequent revision. Many examples from social psychology also support such view (e. g. bystander effect, conformity, blind obedience) [20].

## 3. Conclusions

An agent playing computer games operates between two realities – an Open Reality (external world) and a Closed Reality (virtual world). An agent only accesses Closed Reality via indirect interface (e. g. a keyboard and a mouse, an LCD screen, a speakers, etc.). A Closed Reality has lower complexity than an Open Reality. When an agent enters a Closed Reality his cognitive representation changes from an open to a closed one. This triggers a number of measurable cognitive changes, (e.g. the state of flow). Another aspect is the impact of an avatar on an agent in avatar-based computer games. Interaction between an agent and his avatar in a Closed Reality causes a temporary change in a Cognitive Self. That is why all the cognitive alterations occur immediately and non-inferentially after an agent enters a Closed Reality. An agent is not aware of himself being immersed in activity and neither is he aware of this temporary change in the representation of his Cognitive Self.

## References


[1] Baudrillard, J. (2000) *Simulacra and Simulation.* Michigan: The University of Michigan Press, p. 1-2
[2] Roush, W. (2007) Second Earth. *MIT Technology Review Online*. http://www.technologyreview.com/
[3] *Virtual Game is a 'disease model'.* BBC News Website: http://news.bbc.co.uk/1/hi/health/6951918.stm
[4] Entertainment Software Association: http://www.theesa.com
[5] Chater, N., Vitanyi, P. (2003) Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences* 7, 19-22
[6] Hochberg, J., McAlister, E. (1953) A quantitive approach to figure 'goodness'. *Journal of Experimental Psychology* 46, 361-364.
[7] Chater, N. (1996) Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review* 104, 301-318
[8] Kofta, M., Dolinski, D. (2002) Cognitive approach to personality. W: Strelau, J. *Psychology* 2, 561-600, Gdansk: GWP.
[9] Falk, R., Konold, C. (1997) Making sense of randomness: implicit encoding as a bias for judgment. *Psychological Review* 104, 301-318
[10] Ting-Jui Chou, Chih-Chen Ting. (2003) The role of Flow Experience in Cyber-Game Addiction. *CyberPsychology & Behavior*, 6, 663-675.
[11] Csikszentmihalyi, M. (2002) *Flow*. London: Rider.
[12] Csikszentmihalyi, M. (1972) Beyond boredom and anxiety. San Francisco: Jossey-Bass.
[13] Hoffman, H. G. (2004) Virtual-Reality Therapy. *Scientific American Online*: http://www.sciam.com/
[14] Wood, R. T. A., Griffiths, M. D., Chappell, D., Davies, M. N. O. (2004) The Structural Characteristics of Video Games: A Psycho-Structural Analysis. CyberPsychology & Behavior 7, 1-10.
[15] Waelbers, K., Spence, E. (2007) Ethics in the Virtual World. *Extended Abstract for European Computing and Philosophy Conference in Twente*, Netherlands, Twente.
[16] Strawson, G. (1997) The Self. *Journal of Consciousness Studies*, 4, 5/6, 405-428
[17] Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, 94, 319-340.
[18] Baumeister, Roy. (1991) *Escaping the Self: Alcoholism, Spirituality, Masochism and Other Flights from the burden of Selfhood*. New York: Basic books
[19] Markus, H., Nurius, P. (1996) Possible selves. *American Psychologist*, 41, 954-969.
[20] Braisby, N., Gellatly, A. (2005) *Cognitive Psychology*. Oxford: Oxford University Press

This page intentionally left blank

# Part II

# Living with the Golem: Robots and Autonomous Agents

This page intentionally left blank

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Can a Robot Intentionally Conduct Mutual Communication with Human Beings?

Kayoko ISHII
*National Institute of Science and Technology Policy (NISTEP)*
*Ministry of Education, Culture, Sports, Science and Technology (MEXT)*
*Japan*

**Abstract.** The question of whether a robot can communicate with human beings evokes another question: How can human beings have the feeling that they are usually successful in mutual communication? This question may be answered because the emergence of the mind of individuals and 'the mind of community' are not completely separable. The mind of the community may precede the mind of the individual in society. Complex mechanisms of the emergence of the mind of the community and that of the individual may be effectively studied with Cognitive Robotics in Japan. To promote the study, I develop a hypothesis named "a fabulous game of human beings," in which each individual can guess the contents of her/his mind effectively by reading the attitudes and mind of others.

**Keywords.** Social reference, theory of mind, mirror neurons

## Introduction

In the early 1970s, after the social turbulence and awareness of the negative effects of modern science and technology (S&T), Japan tried to convert its S&T policy to be implemented in a scientific way and create new comprehensive S&T to solve social problems by using S&T. These are still important subjects today. My interest is to integrate diversified disciplines of S&T, the humanities, and social science using S&T to better understand actual human beings and ordinary worlds.

In Japan, robotics researchers are developing robots that can leave the controlled laboratory or factory environment and interact with the general public in society. To date, such robots have been favorably accepted in society. A comprehensive science based on robotics, Cognitive Robotics, can promote the integration of diverse disciplines and ground the integrated knowledge on the real world [1]. In Cognitive Robotics, one of the primary questions would be, "Can a robot intentionally conduct mutual communication with human beings?" Thus far, robots are designed to conform to human beings. It is uncertain, alternatively, if humans could think that they are actually communicating with robots. Thus, the question evokes another question about how human beings can usually lead their social lives thinking that they are managing

mutual communication with others. This new question can be effectively studied with Cognitive Robotics.

To promote the study, I propose a hypothesis that humans have launched a fabulous game where only the contents of the mind of others can be read while the contents of one's own mind have to be guessed. By continuing to play the game, each person comes to feel as if s/he can read the contents of her/his own mind by her/himself.

## 1. Integration of Studies on the Human and Society

In Japan, traditional studies aimed, in principle, to cultivate a person's total character. After traditional studies were widely influenced or replaced by analytical occidental studies in the latter half of 19th century, excessive emphasis on the individual began. This was also the very moment in Europe when natural philosophy was transformed into "science" in the present sense through the marriage of mathematics and experimental philosophies and by approaching experimental philosophy and technologies. Since then, science has become more diversified and has developed into specialized disciplines by increasing in one way the strictness of each domain in which to explain nature and the world under controlled situations. Alternatively, science has been losing the potential to discuss the human being as a whole and the world viewed through the eyes of people who are not scientists.

Recently there has been a clear need to integrate diversified disciplines for better understanding of actual human beings in the real world. Indeed, this integration is becoming possible. An increasing number of researchers are interested in studying uncertain, complex and interdependent phenomena concerning living human beings. Researchers are not only interested in the young healthy adult but also in the fetus, infant, child and elderly. Research interests include their development, change or aging in their environment, community and society.

### *1.1. Cognitive robotics*

From its origins, robotics has been multidisciplinary. In Japan, since around 1994, robotics researchers have organized research groups, such as that for Sociointelligenesis, with the primary aim of elucidating human cognition, development, and behaviors by using robots. In 1995, Japan selected the research theme of 'Creating the Brain' as one of the three targets for Japanese brain science. The framework represents researchers' attempts to understand brain functions through "cycles of creating theories and models of information processing of the brain, their verification through experimental science using robots or computer, and improvement of theories and models."

Cognitive robotics refers to a comprehensive science in which robotics, neuroscience (ranging from the experimental to the theoretical or mathematical variety and neuroinformatics), cognitive science, psychology (psychophysics and behavioral measurement) and behavioral sciences seamlessly collaborate in unity while keeping variations in perspective, closely connecting with fields, such as philosophy, social science, anthropology, and economics; exchanging their knowledge and methodologies; and executing mutual verification [1]. Robotics herein represents an expectation of interdisciplinary integration, rather than a simple collection of

independent research fields, and will be realized by using robots as the common verification platform to highlight weaknesses and errors in research processes in individual disciplines and contradictions among different disciplines.

### *1.2. The mind of the community*

As research on cognition, affections and behaviors of humans advances, a longstanding question for the Japanese has surfaced about whether it is appropriate to study human cognition as the properties of just each individual. Experimental studies on mirror neurons suggest the existence of common neural information processing in perceiving others' actions and expressions of emotions and in evoking, performing, and expressing the same actions and emotions in the self. Other studies suggest the existence of shared neuronal mechanisms for sensory predictions of one's own actions and for predictions of others' actions. These mechanisms enable a seamless coordination of actions done by oneself and by others.

These findings urge us to revise the old model that an individual becomes capable of understanding other individuals by extending understanding of the self, and then understanding society as an extension of relationships between the self and other individuals. The understanding of others and the self may simultaneously develop or indeed understanding others may precede self-understanding when a human infant develops in a given human community. Even an adult may unconsciously refer to a kind of "the mind of the community" before s/he has feeling that s/he decided something independently. In turn, "the mind of the community" is maintained by continuous interactions among members of the community.

### *1.3. Old and new aspects of Japan*

The idea of "the mind of the community" may not seem foreign to the Japanese because the most commonly used Japanese word for human is *ningen* 人間. *Nin* means (a) human being(s), and *gen* means between or among. *Ningen* indicates not just a human being as a biological entity but also (a) human being(s) leading social life (lives) and a system in which they live together in a shared community.

In general, Japanese people very much care how others perceive them. A Japanese author quoted a European joke to illustrate that even Europeans know how much Japanese care what others people think of them. In the joke, people are asked to write a book on elephants, and the Japanese wrote one titled "what elephants think about the Japanese." Some years ago this could have been perceived as humor, but now the perceptions of others can be a serious scientific subject: how people obtain information, such as what others in the society think about them; how they use the information; in what conditions they ignore or neglect the information; and why. For the time being, the Japanese will remain unique people who can seriously discuss 'what robots may think of them.'

Japan has developed state-of-the-art robotics. This is one of the earliest countries entering the postmodern era, but traces of animism can still be seen in Japanese society. It may be so deep-rooted in their mind that animism could be the basis for their tendency to not resist finding emotion in robots although they completely know by reason that robots are mere machines. The Japanese can easily have empathy and project emotions on humanoid robots. It will be worth conducting Cognitive Robotics

on relationships between the mind of the self and that of the community and studying mechanisms by which human beings implicitly integrate the viewpoints of others.

## 2. Mutual Communication in Human Society

The question "Can a robot intentionally conduct mutual communication with human beings?" represents a set of subjects intensely discussed in Cognitive Robotics, such as heuristics, consciousness, intention, autonomy, development, understanding and expressing of affect and emotion and social skills. Instead of directly confronting this question, I deconstruct the question to make the integrative studies in Cognitive Robotics easier.

'A robot' or 'robots as a species'? Human communication abilities have developed as those of a species, not as those of an individual. For example, newborn babies are endowed with innate mechanisms that enable them to acquire human language, but the fact that an infant can indeed begin to speak or which language to speak depends on a system of social interactions into which the infant is integrated.

'Mutual' or 'unilateral'? What can 'mutual communication' between human beings and robots be like? At present, actual robots are designed to conform to humans and often to elicit human sympathy. For the time being, when human beings find intentions, feelings, and other mental properties in robots, this can be attributed to human empathy and projection of emotions. It is not clear, however, if humans can actually accept robots as human companions.

Scientists often expect too much of a robot. To begin, the question is if a human being always succeeds in quasi perfect communication with other members of society. When humans think they can, does such communication consist of completely conscious and intentional processes? It does not seem so. Why can most human beings usually neglect these facts? It may be because the mind of the self has co-evolved and co-developed with the mind of community.

### *2.1. Evolution of information processing in the human*

Humans are social animals. All individuals have to seek a compromise with others even if they cannot reach perfect mutual understanding with each other. It is supposed that huge amounts of information processing in the human brain should be devoted to managing social life. This processing is the default system and the load does not usually attract our attention. Certain people who do not fully develop or lose neural mechanisms responsible for the process help us to notice the existence of the process.

An autistic adult, for example, wrote that it is easy to recognize the mind of the quadruped but difficult for her to read the mind of animals leading complex social lives, such as primates and human beings [2]. The human mind employs complex information processing that cannot be interpreted with the logic of her mental computation.

According to hunter-gatherers, it is easy to read the mind of wild animals but not that of domestic animals because domestic animals are unsettled and rough [3]. Domesticated animals adapt their cognitive styles to that of their oppressor, the human [4]. A group of Japanese anthropologists suggests that humans have modified and are continuing to modify their minds to continue their social lives. The anthropologists call this phenomenon 'self-domestication.' Members of human society may have to learn to

adapt their cognitive styles to that of their oppressor, that is, the other members of society or, rather, the system in which they live together.

In the early stage of human civilization, when an individual encountered strangers from another community, it might not have been the individual her/himself who had to confront the strangers; rather it was "us," that is, family, clan, or tribe.

### *2.2. Social brain*

The average volume of the brain generally increases in proportion to body weight. Primates deviate from this rule, and *Hominidae* especially have a larger neocortex. According to the social brain hypothesis, the human brain has evolved to cope with the complexity within society [5]. There is a correlation between relative neocortical volume and average size of community presented in the number of members. Based on the correlation, a community for *Homo sapiens* is predicted to have around 150 members. This may seem too small for modern citizens, but the number coincides with the average size of a network in the small-world experiment, a traditional clan, an infantry company of the Roman Empire and the modern army. Staff officers know from their experience that this size, around 150 members, is the maximum to maintain a functional community within which members can know each other despite the advance in communication technologies.

What makes human community so complicated? Some of the causes are supposed to be diverse criteria to judge relationships with others before deciding one's own attitudes and actions:

a) hierarchy of social order, physical strength, skills, and other factors
b) union depending on blood relationships

Although both categories are important, one often has to choose which category or which criteria in the category to place priority on, depending on changing situations. The decision should be made in a limited time. Even after a choice, an individual has to consider how to conciliate slighted relatives or neglected fellows. A great difference between human cognition and that of other animals may be the heavy loads of consideration on third parties that are not necessarily present in a specific situation.

### *2.3. An outside to inside model*

One of the earliest pro-communication behaviors that human infants show is "referential looking." When they face novel circumstances or dangers, they spontaneously look at their caregivers, notably their mothers. This may be a physiological response elicited by internal changes such as affections and emotions. Mechanisms underlying "referential looking" may be innate and emerge upon biological scaffolds.

Then "social reference" appears in which a child inquires after her/his caregiver's expression and attitudes and accordingly modifies her/his behaviors (Figure 1). In order that "social reference" develops, infants need caregivers who actively act upon them and interpret what is taking place between them.

"Self reference" comes later than "social reference." An adult reflects in oneself when her/his action is disturbed or suspended and, accordingly, modifies behavior of the self. S/he refers to the internal state of the self. It is unclear if such an internal state to be referred to has already existed or is created for explanation only by the act of referring.

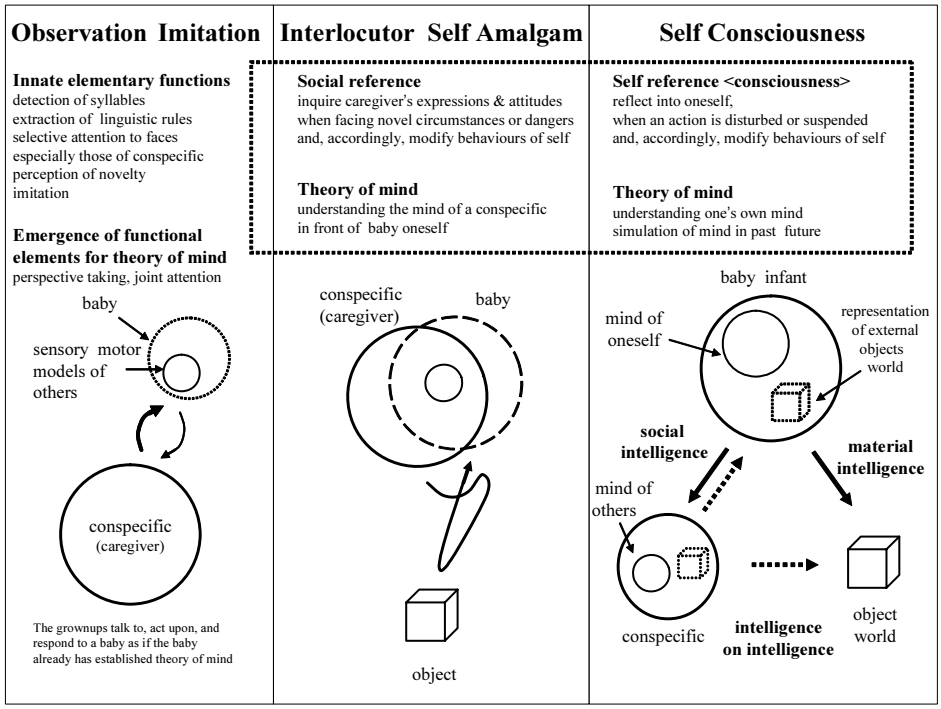



**Figure 1.** From social reference to self reference

Concerning also theory of mind, an infant understands at first the mind of a conspecific in front of her/himself. Then only in the latter stage, s/he understands her/his own mind and can simulate the mind in the past or future.

It is interesting to study if infants convert mechanisms underlying "social reference" to develop "self reference," and if mechanisms underlying theory of mind to understand the mind of others are employed to understand the mind of themselves.

The emergence of "self reference" may be closely related to that of consciousness. According to some Japanese researchers,

- Consciousness can only be shaped against a "background" of unconsciousness. Unconsciousness precedes consciousness either ontogenetically (developmentally) or phylogenetically (evolutionarily) [6].
- Consciousness is a process of becoming aware of any inhibition against thinking or introspection, and a process of introspection in which such awareness elicits past inhibitions against behavior (physical and mental) [7].
- Consciousness is defined as making approximations by performing a highly simplified fictitious series of computations to solve unconsciously generated inconsistent configurations associated with massive and parallel sensorimotor integration [8].

- Consciousness is not the cause of cognition, but only a result. Consciousness is a specific state of working memory and is meant to model unconscious manipulations as simply as possible and to store the results as episode memory [9].

## 3. A Fabulous Game of Human Beings

This raises questions about the appropriateness of studying cognition of *ningen* (human beings) from modern and occidental viewpoints that emphasize such notions as 'individual, benefit/loss, and computation of odds.' Studies in cognitive science suggest that people often cooperate with others simply because they share membership with others without or before strictly calculating if cooperation would be more advantageous. People act in this way for even a single occasion when they cannot expect compensation for their loss in successive trials. This propensity may be distorted in laboratory environments because of prejudice that, in a given artificial game, subjects are expected to defeat other participants and win the game [10]. The requirement is to create a game, which excludes such prejudice.

### *3.1. A fabulous game of human beings*

Thus I propose the hypothesis that human beings launched in the early stage of their evolution "a fabulous game" to co-develop fictitious states of the mind of the self and of the community. Even today newborn babies have to learn the game and, in society, their caregivers spontaneously try to provide active teaching of the game by pretending to play it together. Even adults continue playing the game but usually without consciously noticing they are doing so. This may be tested in Cognitive Robotics by conducting experiments on robots and human subjects, and integrating findings from history, anthropology, social sciences, cultural studies, and other fields.

The rules of the game are set as follows:

1. Early stage
   - Contents of the mind of each person are presented on a card.
   - S/he cannot read her/his own card, but can to a certain extent read the other person's card.
   - By observing another person's reactions to her/his card and subsequent attitudes or behaviors, s/he guesses the contents of her/his card {intercourse of direct interlocutors (individuals A & B) in Figure 2}.
   - One who correctly guesses the contents of her/his own card earlier will be highly appreciated in the community.

2. Chronic stage
   - After members of the society have continuously played this game for a long time, each member begins to feel that s/he can read her/his card by her/himself. This feeling changes to another feeling that 's/he can read her/his mind by her/himself,' and then that 's/he knows her/his own mind.'

- In such situations, a person (A) may expect that if s/he dispatches the contents of the card by using certain symbols (message) to someone else, the recipient (C in Figure 2) will understand her/his mind.
- It is always uncertain if the recipient perceives the message as the sender expected: (misunderstanding).
- Within human society, members inevitably pay attention to third (fourth, fifth, sixth) parties. Consequently, in the game too, a player takes others into account, considering who may notice her/his exchange with her/his interlocutor. S/he cares a great deal how others think of her/his behaviors. This should be an important factor for the specialization of human cognition
- A player who could make a greater number of more influential members of society think that s/he has more valuable cards than an interlocutor is highly appreciated in society.

When the game is deeply integrated in society, a member sometimes gets the feeling that s/he reads the mind even if there are no representations on the card.

### *3.2. Constructivist approaches*

The hypothesis of "a fabulous game of human beings" would be suitable for study by the constructivist approach. The approach is required to study complex phenomena, which analytical approaches or reductionist methods can hardly start to work on or entirely elucidate. This scrap and build approach may have been a favorite way of Japanese researchers. In addition, the Japanese government also appreciates curiosity driven basic research, which requires a long time.

When a theme of research is closely related to complex phenomena of real human beings and society, it is better to leave the controlled laboratory environments and examine the subject in the actual environment. Thus far scholars in the humanities and social sciences have presumed that they can remotely analyze and describe the subjects of their research in controlled environments but disliked intervening into the subjects in the actual environment. Recently, we have been convinced that studies of complex phenomena can be facilitated or accelerated by knowing the effects of interventions on subjects. Sometimes only this kind of approach is available or feasible. Of course, researchers are obligated to make every possible effort to refrain from giving subjects undesired negative effects.

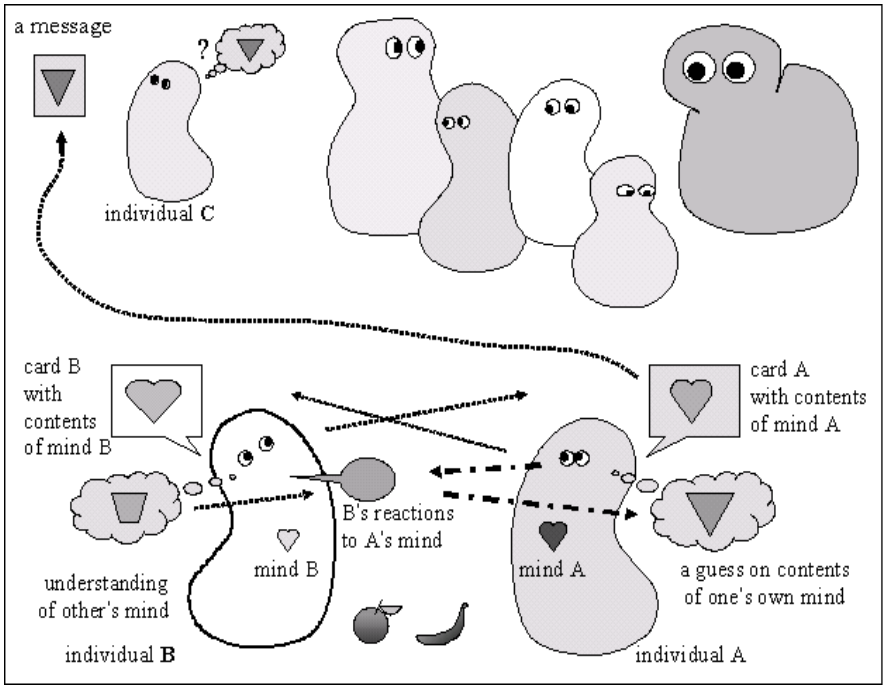



**Figure 2.** A fabulous game of human beings

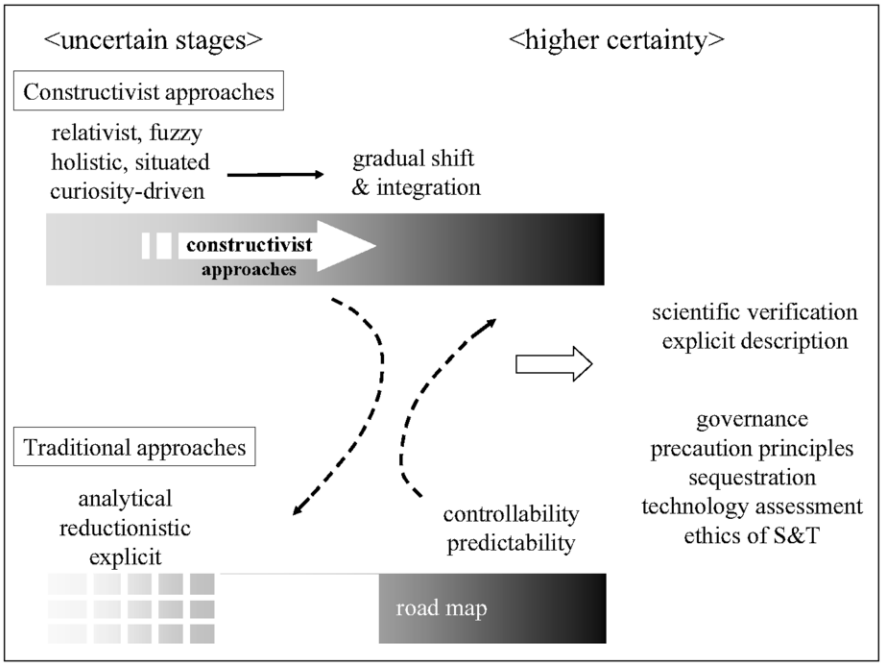



**Figure 3.** Constructivist approaches

## 4. Conclusion

It is true that modern law is based on the existence of free will and the concept of personal responsibility. Research in neuroscience, cognitive sciences, psychodynamics, and other fields presents one piece of inconvenient evidence after another against such a notion of the individual. The human brain and its functions may largely emerge from biological and physical scaffolds and by their interactions with the environment. Contents of the individual mind may be molded in accordance with the mind of the community, which was shaped as a summation of individual thoughts.

I proposed a hypothesis named "a fabulous game of human beings" to be studied in Cognitive Robotics before directly addressing the question "Can a robot intentionally conduct mutual communication with human beings?" We do not know if there could be perfect communication. We cannot even state that a human being always succeeds in quasi perfect communication with other human beings. What is important is to ponder why human beings can compromise and how to find a point of compromise. An important clue may be the fictitious mind of society, and such seems to be increasingly important in the postmodern era.

As humanoids become more widespread and more humanlike, even the Japanese public's favorable attitudes toward robots could sour. It is possible, however, that Japan may find an original way to develop robots based on the traditional Japanese emphasis on cooperation and avoid problems of robot uncontrollability. Such robots could be used for life outside the laboratory. If this could be realized, robotics would develop as a unique area of research.

## References

[1] K. Ishii, Cognitive Robotics to Understand Human Beings, *Science & Technology Trends Quarterly Review 20,* 2006, 11-32. http://www.nistep.go.jp/achiev/ftx/eng/stfc/stt020e/qr20pdf/STTqr2001.pdf
[2] T. Grandin, Thinking in pictures and other reports from my life with autism, Vintage Books, New York, 1995.
[3] H. Brody, The other side of Eden, Hunter gathers, Farmers and the shaping of the world, Douglas & McIntyre, Canada, 2000.
[4] K. Omoto, Self-domestication of human beings and the modern period. Jinbun-shoin, Tokyo, 2002.
[5] R.I.M. Dunbar, Grooming, Gossip and the evolution of language, Faber & Faber, London, 1996.
[6] S. Shimojo, What is Consciousness? Kodan-sha, Tokyo, 1999.
[7] T. Kitamura, Can a Robot Have a Mind? - Introduction to Cyber Consciousness Theory. Kyoritsu Shuppan, Tokyo 2000.
[8] M. Kawato, The Computational theory of the brain. Sangyo-Tosho, Tokyo, 1996.
[9] T. Maeno, How to Make a Conscious Robot - Fundamental idea based on passive consciousness model, *Journal of the Robotics Society of Japan*, 23, 2005, 51-62.
[10] A. Ito, M. Ohashi, T. Itatsu, & K Terada, Why it is so difficult for humans to solve non-zero-sum games, *Proceedings of The 24th Annual Conference of the Japanese Cognitive Science Society,* 2007, 264-269

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# On the Ethical Quandaries of a Practicing Roboticist: A First-Hand Look

Ronald C. ARKIN
*Mobile Robot Laboratory*
*College of Computing*
*Georgia Institute of Technology*
*Atlanta, GA 30332, U.S.A.*

**Abstract**. Robotics has progressed substantially over the last 20 years, moving from simple proof-of-concept experimental research to developing market and military technologies that have significant ethical consequences. This paper provides the reflections of a roboticist on current research directions within the field and the social implications associated with its conduct.

**Keywords.** Robot ethics, military robotics, entertainment robotics

## Introduction

I have been a practitioner in the field of robotics for over 20 years, and during that time I developed a strong appreciation for the potential ramifications of the research that I have been and currently am conducting, ranging from the purely scientific to the more applied. This has led me to delve deeply into the questions surrounding the ethical practice of robotics as a whole and to seek out the means for analysis of the consequences of my personal actions in the field (past, present, and future) while also actively encouraging my colleagues to do so.

There are all sorts of red flags being raised by others regarding the perils of robotics, all the way from a predicted end of the human-dominated world due to self-replicating robots (e.g., [1,2]) to far more immediate issues surrounding the application of robotics (e.g., the use of robots in warfare [3-5], labor ramifications, and the deliberate psychological manipulation of human beings by robot entities [6-8]). While I could also take a stand on the more alarmist perspectives, I will in this article, address those concerns arising from the here-and-now practice of robotics from a personal perspective, most of which have serious short-term ethical consequences. While some of these issues have been discussed in prior Roboethics conferences (e.g., [9,10]) in a more general informative manner, they have not been developed in the context of an individual researcher's perspective nor, oddly enough, in a true ethical context, where

different theories of ethical reasoning are applied, whether they be utilitarian, cultural relativism, social contract theory, Kantian, etc.

Independent of the specific personal ethical framework for analysis chosen, I will lay in front of you three ethical quandaries that are not hypothetical but constitute the reality that I have been or am currently confronted with. In teaching my class on robot ethics (CS 4002 Robots and Society) I encourage my students to examine not only abstract or removed case studies but also current practices such as my own in light of the criticisms they may well be subjected to within society. I find this exercise invaluable personally as well, as it informs me, often in surprising ways regarding the views that at least one segment of the population holds regarding this research.

## 1. First Quandary: Autonomous Robots Capable of Lethal Action

One major research area I am responsible for involves military robotics. While I choose to only conduct unclassified research so that I can publish and talk freely about my work (at least to date), my experience in this area ranges from areas such as robots for explosive ordnance disposal and humanitarian demining to the development of software for autonomous weapons-bearing unmanned vehicle systems (e.g., the Defense Advanced Research Project Agency's (DARPA) Unmanned Ground Combat Program). The controversy surrounding this application is clearly evident, ranging from the traditional arguments against warfare in general and new weapon construction in particular, to issues surrounding the direct application of lethality by autonomous systems without having a human in direct control or issuing a confirmation of an order to kill. Ongoing research on my part for the U.S. Army involves assaying opinion (of the public, researchers, the military, and policymakers) on the use of this latter class of autonomous robots, while also investigating how to embed an "artificial conscience" in these vehicles to ensure that the international laws of war and rules of engagement are strictly followed by machines, perhaps even more effectively than by humans. This has required developing an understanding of Just War theory [11] and delineating methods by which combatant/noncombatant discrimination, proportionality of force, minimization of collateral damage and unnecessary suffering, and related *Jus in Bello* ethical issues can be enforced within autonomous robots. This research speaks predominantly to the deontological basis as encoded in International Conventions (e.g., Hague and Geneva Protocols) in addition to utilitarian considerations in terms of military necessity, weapon selection, firing pattern, and permission to fire, all subject to the former rights-based restrictions.

As it is clear to me that the technology that I helped create, specifically autonomous robotic architectures, is moving forward in warfare applications, with or without my participation, I feel compelled to act in a manner that leads to the development of autonomous systems that are capable of complying with accepted International Law. Further I have become convinced of the highly controversial position, due to the propensity of soldiers to tolerate and commit illegal acts under wartime conditions [12], that ultimately autonomous systems can outperform humans ethically in the battlefield. Details of this ethics-driven approach can be found in [5].

## 2. Second Quandary: Entertainment Robotics and the Suspension of Reality

The second area of ethical controversy deals with personal robotics. I have served as a consultant for Sony Corporation for nearly 10 years in the development of software for the AIBO and QRIO entertainment robots (Figure 1) [13]. While most researchers view this activity as an innocuous and even beneficial use of robotics, possibly for the treatment of isolated elderly people, not all agree [6,7]. This research requires a deep understanding of not only a robot's capabilities but also human psychology, where the roboticist's deliberate goal is to induce pleasant psychological states in the observer through specialized patterns of robot behavior and, to the greatest extent possible, suspend observer disbelief that this robot is not alive. The intended goal is to establish a long-term, even lifelong, human-robot relationship, perhaps even a dependency not unlike what is experienced with pets or among friends.

Some view the ethics for this type of research as no different than that of advertising, cinema, video games, or other forms of entertainment. Others such as Sparrow [6] argue that this is an intrusion into the rights of the elderly to remain in contact with the real world, while society (and researchers such as myself) makes excuses for its intended unethical use. Robotics researchers make a tacit assumption that the creation of this new technology is wholly appropriate and can only enrich the lives of those on the receiving end. It is important that we as scientists re-examine this assumption. To that end, I have gone to the heart of the scientific source [14] to present contradictory perspectives based on deontological arguments and potential violations of the social contract that challenge the underlying assumption that the goal of this form of robotics research is ethically acceptable.

**Figure 1.** QRIO is the humanoid on the left; AIBO is the robot dog on the right

## 3. Third Quandary: Robotics and Unemployment

The final area of personal concern involves the displacement of workers, in areas such as shipyards. Although I currently have limited current research in this area, I am considering and willing to expand it, which in many respects has caused me more soul-searching than the other two examples cited above. I have also conducted research extensively in the applications of autonomous robotics to manufacturing in years past [15,16].

Indeed much of the underlying premise for the use of robotics as a whole is the elimination of the three D jobs: those that are Dull, Dangerous, and Dirty. While this at first blush appears to be a noble goal, without concomitant social support we are just encouraging the same forms of social upheaval that accompanied the earlier industrial revolution. From a corporate perspective, this research avenue can undoubtedly lead to a clash between an act utilitarian perspective of a large industrial concern with the individual worker's (Kantian) right to good will.

When a roboticist can project the consequences of their research as ultimately leading to significant unemployment with worldwide impact, and while being unable to directly influence social support structures for those potentially made unemployed, what is their moral responsibility here? This may lead to a more traditional debate on industrial revolutions in general, but nonetheless roboticists often are woefully unaware of where the consequences of their work may lead in this domain.

## 4. In Summary

These issues are personal day-to-day concerns, and I contend they should also be part of a regular professional roboticist's diet. As in many ethical areas, we will not agree universally on the outcomes for these and other related issues, at least from an individual perspective. Nonetheless, I argue that it is a central responsibility of a roboticist to conduct such self-examinations to ensure that he/she is aware, at least consistent with their own morality, the consequences of their actions and also to be prepared to become engaged with others in this field on related ethical concerns, so that we as a group of concerned scientists can develop acceptable limits and guidelines to a broad range of emerging robotics issues. Reaching out and engaging others from non-technical communities such as philosophers, social and political scientists is crucial toward achieving this end.

## Acknowledgments

A portion of this research is funded under Contract #W911NF-06-0252 from the U.S. Army Research Office.

## References

[1] Joy, William, "Why the Future Doesn't Need Us", Wired, Issue 8.04, April 2000.
[2] Moravec, Hans, Mind Children: The Future of Robot and Human Intelligence, Harvard University Press, 1990.

[3] Sparrow, R., "Killer Robots", Journal of Applied Philosophy, Vol. 24, No.1, 2006.
[4] Asaro, P., "How Just Could a Robot War Be?" presentation at 5th European Computing and Philosophy Conference, Twente, NL June 2007.
[5] Arkin, R.C., "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture", Technical Report GIT-GVU-07-11, College of Computing, Georgia Institute of Technology, 2007.
[6] Sparrow, Robert, "The March of the Robot Dogs", Ethics and Information Technology, Vol. 4 No. 4, 2002, pp. 305-318.
[7] Sparrow, R. and Sparrow, L., "In the Hands of Machines? The Future of Aged Care", Mind and Machines, Vol. 16, pp. 141-161, 2006.
[8] Krahling, M., "In Between Companion and Cyborg: The Double Diffracted Being Elsewhere of a Robodog", International Review of Information Ethics, Vol. 6, pp. 69-77, December 2006.
[9] First International Symposium on Roboethics, Villa Nobel, San Remo Italy, January 2004. http://www.roboethics.org/sanremo04/ROBOETHICS_Program.html.
[10] Proceedings of Roboethics Workshop at 2007 IEEE International Conference on Robotics and Automation, Rome, Italy, April 2007. http://www.roboethics.org/icra07/contributions.html.
[11] Walzer, M., Just and Unjust Wars, 4th Ed., Basic Books, 1977.
[12] Surgeon General's Office, Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05-07, Final Report, Nov. 17, 2006.
[13] Arkin, R., Fujita, M., Takagi, T., and Hasegawa, R., "An Ethological and Emotional Basis for Human-Robot Interaction", Robotics and Autonomous Systems, 42 (3-4), March 2003.
[14] Arkin, R.C., "Ethical Issues Surrounding the Use of Robotic Companions for the Elderly: Illusion versus Reality", presentation at Workshop on Assistive Technologies: Rehabilitation and Assistive Robotics, held at IEEE/RSJ 2007 International Conference on Intelligent Robotics and Systems (IROS '07), San Diego, CA, October, 2007.
[15] Arkin, R.C. and Murphy, R.R., "Autonomous Navigation in a Manufacturing Environment", IEEE Transactions on Robotics and Automation, Vol. 6, No. 4, pp. 445-454, August 1990.
[16] Murphy, R. and Arkin, R.C., "Autonomous Mobile Robots in Flexible Manufacturing Systems", Proc. Fourth International Conference on Artificial Intelligence Applications, San Diego, CA, pp. 412-414, 1988.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*


# How Just Could a Robot War Be?

Peter M. ASARO
*HUMlab & Department of Philosophy, Umeå University*
*Center for Cultural Analysis, Rutgers University*
*peterasaro@sbcglobal.net*

**Abstract.** While modern states may never cease to wage war against one another, they have recognized moral restrictions on how they conduct those wars. These "rules of war" serve several important functions in regulating the organization and behavior of military forces, and shape political debates, negotiations, and public perception. While the world has become somewhat accustomed to the increasing technological sophistication of warfare, it now stands at the verge of a new kind of escalating technology–autonomous robotic soldiers–and with them new pressures to revise the rules of war to accommodate them. This paper will consider the fundamental issues of justice involved in the application of autonomous and semi-autonomous robots in warfare. It begins with a review of just war theory, as articulated by Michael Walzer [1], and considers how robots might fit into the general framework it provides. In so doing it considers how robots, "smart" bombs, and other autonomous technologies might challenge the principles of just war theory, and how international law might be designed to regulate them. I conclude that deep contradictions arise in the principles intended to govern warfare and our intuitions regarding the application of autonomous technologies to war fighting.

**Keywords.** Just war theory, robots, autonomous systems

## Introduction

Just war theory is a broadly accepted theoretical framework for regulating conduct in war, that has been embraced by such esteemed and influential institutions as academia, the US military establishment (including the military academies[1]), and the Catholic Church. It is also compatible with, if not actually a formulation of, the principles underlying most of the international laws regulating warfare, such as the Geneva and Hague Conventions.

[1] Walzer's book was a standard text at the West Point Military Academy for many years, though it was recently removed from the required reading list.

This paper aims to illuminate the challenges to just war theory posed by autonomous technologies. It follows Michael Walzer's [1] articulation of the theory, which has been the most influential modern text on just war theory. While there are compelling criticisms of Walzer's formulation (*e.g.* [2]), it is his articulation which has had the most influence on the institutions and international laws regulating war.

Before we begin, I should clarify what I mean by robots and other autonomous systems. "Autonomy" is a rather contentious concept, and its relation to material technologies adds further complications. It is thus useful to think about a continuum of autonomy along which various technologies fall depending upon their specific capabilities. Most generally, any system with the capability to sense, decide and act without human intervention has a degree of autonomy. This includes simple systems, such as a landmine that "decides" to explode when it senses pressure. Obviously, systems with only the most rudimentary forms of sensors, decision processes and actions lack various aspects of full autonomy. The landmine does not decide where it will be placed, and its physical placement largely determines the consequences of its actions, thus it has much less "autonomy" than systems with more sophisticated means of sensing, deciding and acting. If we were to consider it as a moral agent, we would not be inclined to hold it morally responsible for its actions, but rather hold responsible those who placed and armed it. It thus occupies an endpoint in the continuum of autonomy and moral responsibility.

Certain kinds of "precision" weapons (*e.g.,* "smart" bombs) use global-positioning systems (GPS) and sophisticated control mechanisms to deliver them accurately to a target. The selection of a target, and the determination of its location, value and risks, is still determined by human agents who control the weapons system, however. Thus we might wish to "blame" a smart bomb, or its design, for failing to reach a designated target, but not for the selection of the target. It thus represents a point further along this continuum, and shares this position with various kinds of guided weapons and automated anti-aircraft batteries (*e.g.* Patriot missile systems, and the Phalanx gun systems) and automatic anti-ballistic missile systems (*e.g.* Star Wars/SDI), that detect and destroy sensed threats without immediate human intervention. Though, again, such an armed system is still dependent on responsible human decisions as to when it is appropriately activated.

Still more autonomous are systems which use sophisticated sensor analysis to select appropriate targets on their own and make decisions about the appropriateness of various actions in response to its situation. The emerging technologies of robotic weapons platforms incorporate some or all of these features, using image processing to identify targets, and selecting from a broad range of offensive and defensive actions in their engagement of targets. These are technological capabilities which already exist, and are beginning to be implemented in various countries. These systems tend to be designed to seek permission from human authorities before using lethal force against a target, what the US military calls the "human-in-the-loop," but this is not a technological necessity. We can identify the choice to use deadly force against a specific target as a critical threshold along the continuum of autonomy, and one which carries a greater moral burden in the design and use of such a technology. There are, however, systems with more autonomy than this.

As robotic technologies advance, it is possible that they will acquire moral capacities that imitate or replicate human moral capacities. While some systems might merely enact pre-programmed moral rules or principles, autonomous robotic agents might be capable of formulating their own moral principles, duties, and reasons, and

thus make their own moral choices in the fullest sense of moral autonomy. There are many possibilities short of replicating a fully autonomous moral subject, such as agents with moral awareness but not the freedom of choice to act upon that awareness. While this still remains in the realm of science fiction, it does not seem impossible in principle that a robot could achieve autonomy in a Kantian sense, in which it takes responsibility for its actions, reasons about them morally, and identifies itself with the moral quality of its own actions. At some point along the continuum, but probably before Kantian autonomy, various questions will arise about the moral responsibilities of others towards such autonomous systems, and in particular whether these systems have moral rights.

There are many degrees of autonomy along the continuum at which specific systems might fall, so I will consider the implications of these on the interpretation and application of just war theory to various situations. I should also say something at this point about the speculative nature of this work. Many people find the idea of robotic soldiers and robotic moral agents to be somewhat fantastical, the stuff of science fiction and not something that deserves serious consideration. My arguments are meant to cover technological possibilities that do not yet exist, and even some that shall perhaps never exist, yet I believe that it is important to develop our understanding of existing technologies in light of these hypothetical possibilities. There is a very great interest in building autonomous systems of increasing complexity, and a great deal of money is being invested towards this goal. It does not seem to be an unreasonable prediction that within the next decade we will see something very much like a robotic soldier being used.[2] If we consider the amount of time and effort it took moral and legal theorists to come to terms with the atomic bomb, then it makes sense to start thinking about military robots now, before they appear fully formed on the battlefield. Robots may not have the same potential to reshape global politics that the atomic bomb did, though indeed they may. Still, it is not unreasonable to expect that these technologies might find their way into other security applications, such as policing civilian populations. There is thus a definite necessity, and a certain urgency to establishing the moral framework in which we might judge the various applications of such technologies, as well as the ethics of designing and building them. This examination of just war theory is a part of that broader investigation.

As my concern in this analysis is with the general capabilities of technologies, primarily their ability to act autonomously, and not with the specific technologies used, I will not spend much time discussing how these technologies might work, apart from their degree of autonomy. One might also object that there already exist legal restrictions on the use of autonomous systems in combat, and these have succeeded thus far in keeping humans-in-the-loop of the most advanced military systems being developed, at least in the US, and thus there is no need for such an analysis. While it is true that humans are being kept "in the loop," it is not clear to what extent this is only a contingent truth, and whether this restriction can resist pressures to extend the autonomy granted to military systems [3]. It thus seems reasonable to review the fundamental principles which underwrite existing prohibitions on the use of autonomous technologies, as well as to try to anticipate how they may need to be

[2] The military of South Korea already has plans to deploy autonomous robots armed with machine guns and live ammunition along the border with North Korea. The system is designed by Samsung, and will shoot at any human attempting to cross the DMZ.

augmented or extended to address new technological possibilities as they begin to appear on the horizon.

## 1. Just War Theory

Walzer's [1] just war theory aims to provide a theoretical framework for debate about the morality of specific choices and actions with regard to war by establishing a small set of principles that effectively capture general moral sentiments. Rather than dismiss all war as immoral, it seeks to carefully distinguish those specific acts that are moral so as to deny authority to those who would abuse moral sentiments in taking a nation to war, or try to legitimate immoral acts during the conduct of a war. It establishes a rational framework for distinguishing just from unjust acts of, and in, war and takes a liberal approach which seeks to protect the rights of individuals and states from unjust harms. It derives its rational principles from reflecting on shared moral sentiments and intuitions, and from recognizing the conventional nature of war and the rules of war which tend to govern it. As McMahan [2] makes clear, there are numerous inconsistencies in how Walzer articulates the foundations of the theory in terms of individual rights, state rights, moral sentiments and conventional norms. McMahan's critique, however, aims to preserve most of the overall structure of just war theory by putting it on firmer foundations based in individual rights.

A key distinction in just war theory is drawn between what are just reasons for going to war, *jus ad bellum*, and what are just acts in the fighting of war, *jus in bello*. For Walzer, the two are completely independent of one another. Thus, for Walzer, the actions of soldiers on both sides of a war can be just if they observe the overriding principles of *discrimination* and *proportionality*. The substantive critiques of Walzer challenge this independence, and I believe that, as a moral question, proportionality depends in important ways upon the reasons for going to war. Despite this, the distinction is highly relevant to the practice and legal regulation of war and reflects two distinct moral questions, even if the moral character of how a war is fought cannot be fully determined without considering the reasons for fighting.

## 2. Autonomous Technology and jus ad bellum

There are at least two significant ways in which autonomous technologies might influence the choice of a nation to go to war. The first of these is that such systems could directly threaten the sovereignty of a nation. As such it would challenge just war theory, or the theory might have to be extended to cover this possibility. The second way is one of the reasons many people fear the development of autonomous killing machines, like robotic soldiers, though it is not really a problem for just war theory itself. This is the belief that these technologies will make it easier for leaders who wish to start a war to actually start one, in short that autonomous technologies would lower the *barrier to entry* to war, or be directly responsible for starting a war intentionally or accidentally.

### *2.1. Autonomous technologies challenging sovereignty*

It seems that autonomous technologies offer a great potential for starting wars accidentally, and perhaps even doing so for their own purposes. While this latter potential seems more like science fiction than a real possibility, it is nonetheless useful to consider how just war theory can help aid our thinking about such a situation. The theory accepts only a few very specific causes for war as being just. In principle, only aggression against a nation's sovereignty by another nation is a just cause. Strictly speaking, the aggressor has acted unjustly and the defender has a right to self-defense, and thus may justly fight against the aggression—though history is rarely so clear cut. By extension of this principle, other third party nations may justly join in the fight on the side of nation defending itself against the aggressor, though they are not necessarily *required* to do so, *e.g.* if doing so would threaten their own existence or sovereignty. There are only a few and highly circumscribed exceptions to this principle, those being a pre-emptive strike against an immediately impending aggression, and a humanitarian intervention to stop severe human rights abuses or genocide.

With these general principles firmly in mind, we can imagine new ways in which autonomous technologies could impact upon sovereignty. First, there is the case in which an autonomous technology "accidentally" starts a war. This could be the result of human manipulation, a genuine technical error, or perhaps even by the purposeful intention of the technology. It could also turn against the nation that created it, resulting in one sort or another of a "robot revolution."

#### *2.1.1. Accidental war*

The idea of an "accidental" war is closely related to our conception of the sovereignty of states, though it is not truly a threat to sovereignty. An "act of war" is considered to be an intentional act committed by one state against another. Thus, an unintentional act which is interpreted as an act of war could lead to an accidental war. The possibility of an accidental war has always existed, and generally the decisions to go to war are based on intentions that pre-exist any specific act of war, which is only the proximate cause or a token of justification. Autonomous technological systems introduce new dangers, however, in that they might act in unanticipated ways that are interpreted as acts of war.

There was a common fear throughout the Cold War era that the complex technological control systems for nuclear arms might malfunction, unintentionally starting a nuclear war that nobody could stop. To the extent that all large complex technological systems are prone to unpredictable errors in unforeseeable circumstances, the systems of control for autonomous robotic armies will be too. Further, to the extent that robots are used to patrol dangerous areas, contentious borders and political hot spots, it seems quite likely that some of their actions might be interpreted as acts of war, though no political or military official specifically orders such an act. While this is a real danger, it is not altogether unlike the threat posed by rogue officers and soldiers who accidentally or purposely commit such acts despite not being properly authorized by their chain of command, though they are likely to be aware of the potential for misinterpretation.

The wars that might result from such accidents cannot be just from the perspective of the unintentional aggressor, who then has an obligation to stand down from that aggression. While the state that is harmed by such unintentional acts has a genuine grievance, and has a right to defend itself, if it declares war in response, it will not be a

just war unless the aggressor continues to pursue a war with further acts of aggression. Often, however, groups within one or both states are interested in having a war and will seize upon such incidents as opportunities to escalate hostilities and justify a full-scale war. In any case, it seems that just war theory provides the means necessary to interpret and consider cases of unintentional acts of war, in which both the human chain of command and the autonomous technology do not intend the act of war in the relevant sense.

More complex are cases in which the autonomous technologies have intentions of their own—when they are near the Kantian moral-agent end of the autonomy spectrum. Again, such agents may act unintentionally and the situation would not be unlike those in which human acts unintentionally. However, a new kind of problem arises when autonomous technologies begin to act on their own intentions and against the intentions of the states that design and use them. These situations present many problems. First, it may be difficult to distinguish a genuine intention from a technical error, which casts doubt on just what the intention behind the act is. Further, such a display of incongruent intention might indicate that the autonomous system is no longer under the control of the state, or individual, which produced or employed it. As such, it might not be appropriate to attribute its actions as being representative of the state, *i.e.* it is a rogue agent. It is not clear what responsibility a state has for the actions of rogue agents that it creates or supports, *i.e.* whether it is liable to be attacked. The rogue agents themselves are capable of taking some of the responsibility, however, in virtue of their moral autonomy and are certainly liable to be attacked.

These possibilities also bring into focus a new kind of question, namely whether it is wise, or just, to build and install such automated systems in the first place, given the kinds of risks they engender. This question has been asked most pointedly with automated nuclear defense systems. Here the stakes are very high, and it seems morally wrong to leave the ultimate decision to launch a nuclear assault up to an automatic process rather than a human, who might quite reasonably fail to act out of a moral sense of duty, and thus avert a total nuclear war.[3] In fact, we might want to design military robots in a way that allows them to refuse orders that they deem to be illegal, unjust or immoral, though researchers are only beginning to think about how we might do that.[4] To the extent that autonomous systems begin acting on their own intentions, however, then we might be concerned about their acts of aggression towards their own state, as well as towards other states.

### *2.1.2. Robot revolutions*

The notion of a robot revolution is as old as the stage play in which the word "robot" was first coined, Capek's *R.U.R.* [4], in which worker robots all over the world join in a global revolution and overthrow the humans. While this may seem like a fanciful bit of science fiction, we can ask serious questions about the moral status of such revolutions

[3] There is a rich literature on the morality of "Doomsday" devices in nuclear deterrence, see [5] on this. The deterrence literature is preoccupied with strategic considerations, and with the credibility of the deterrent threat, and thus sees the amoral automaticity as a strategic advantage. Morally, this seems contemptible because it provides for no moral reflection of the consequences of such an action. If we take the notion of autonomous moral agents seriously, then perhaps a machine could make a moral determination without human intervention, something not considered during the Cold War.

[4] The roboticist Ronald Arkin [6] of Georgia Tech is working on just such a robot control architecture, under contract from the US Army.

according to just war theory. Let us imagine a situation in which a nation is taken over by robots—a sort of revolution or civil war. Would a third party nation have a just cause for interceding to prevent this?

In such cases just war theory might be of two minds, depending upon the moral status and autonomy of the robots in question. On the one hand, it is a violation of sovereignty to interfere in the civil war of another nation. On the other hand, it is legitimate to intervene to aid a state that is being threatened by a foreign power. Thus a great deal turns upon whether the robots who lead the revolution are seen as autonomous moral agents with a political right to revolt, or if they are the non-autonomous agents acting on behalf of some other morally autonomous agent, or if they are a non-moral or amoral system under the control of no autonomous agents but simply threaten a state. A further possibility is that the robots might represent a humanitarian crisis if they were seeking to completely eradicate or enslave the humans, implying that humanitarian intervention would be just in such a case.

Even if we consider robotic moral agents as being different than humans in important ways, the question regarding their right to revolution ultimately turns on whether they are entitled to a right to self-determination. For humans this right is tied up in other rights to individual freedom, to not be killed, to not be tortured, *etc.* While it remains to be seen if these rights are separable, due to different technological designs which only partially replicate human mental and moral capacities, we can answer this question based on the relevant portion of those rights. Assuming that we have a theory of what is required for someone to be entitled to a right to self-determination, *e.g.* a sufficiently sophisticated moral autonomy, then the robot rebels will either have this right or they will not. If they do, then just war theory will treat them just as it treats human rebels seeking to take control of their own country. If they do not, then the theory will treat them as agents of another power. If that power has legitimate claims to revolution, then there is no right for third parties to intervene. If they are agents of a foreign or otherwise illegitimate power, or of no autonomous moral agency at all—a sort of man-made disaster—then the threat they pose will justify intervention by outside powers who seek to protect the state from the robots. If in any of these cases the robots represent a humanitarian threat, then third party nations may also intervene on humanitarian grounds.

Thus we can see again that just war theory is able to deal with many of the more fanciful cases in a fairly traditional way. It leaves open, however, the critical question of whether or when machines might be due some or all of the rights of humans, as this is completely outside the scope of theory of just war. It is also not clear that it can always be easily determined whether machines are autonomous and, if not, on whose commands or intentions they are acting.

### *2.2. Technologically lowering the barriers of entry to war*

I believe that one of the strongest moral aversions to the development of robotic soldiers stems from the fear that they will make it easier for leaders to take an unwilling nation into war. This is readily apparent in light of recent history, including the 1991 Persian Gulf War, the 1999 war in Kosovo, and the 2003 invasion of Iraq. These events have brought into sharp relief the complex relationships between the political requirements on national leaders, the imagery and rhetoric of propaganda and the mass media, and the general will of citizens in the processes of deciding when democratic nations will go to war.

Irrespective of the underlying justness of the motives, when the leadership of a state decides to go to war, there is a significant propaganda effort. This effort is of particular importance when it is a democratic nation, and its citizens disagree about whether a war is worth fighting, and there are significant political costs to a leader for going against popular sentiments. A central element of war propaganda is the estimation of the cost of war in terms of the lives of its citizens, even if that is limited to soldiers, and even if those soldiers are volunteers. A political strategy has evolved in response to this, which is to limit military involvement to relatively "safe" forms of fighting in order to limit casualties, and to invest in technologies that promise to lower the risks and increase the lethal effectiveness of their military.

We can see these motivations at work in the NATO involvement in Kosovo in 1999, in which NATO limited its military operations to air strikes.[5] The political pressure to avoid casualties among a nation's own soldiers is thus often translated into casualties among innocent civilians, despite this being fundamentally unjust. Technologies can shift risks away from a nation's own soldiers and can thereby add political leverage in both domestic politics, through propaganda, and in diplomatic efforts to build alliances among nations. Thus, the technology functions not only in the war itself, but in the propaganda, debate and diplomacy that brings a nation into war. In this regard, it is primarily the ability of the technology to limit risks to the nation that possesses it, and its allies, that allows it to function in this way. Certainly the replacement of soldiers by robots could achieve this in a new and somewhat spectacular way, perhaps by eliminating the need for any soldiers from that nation to actually go to the battle zone.

Given these facts, it is quite reasonable to conclude that the introduction of any technology that can limit the risks to a nation's soldiers and civilians would serve a similar function. In some sense, all military technologies that work well serve this function, to some degree or when taken altogether, whether it is better airplanes, or better body armor, or better bombs, or better communications, or better strategies, or better robots, or even better battlefield surgery techniques. Indeed, recent US media has spent a great deal of energy trumpeting the technological sophistication of the US military in this way. The ultimate aim of all military technology is to give an advantage to one's own soldiers, and this means limiting their risks while making it easier to kill enemy soldiers and win the war. So in general, all military technological development aims at this same objective. Moreover, even with the most sophisticated machinery, and guarantees of extremely low casualties, most citizens in most countries are still averse to starting an avoidable war, and are nearly always averse to starting an unjust war.

---

[5] It is important to note that Walzer, in his introduction to the third edition of [1] in 1999, has criticized this particular decision as being unjust because of the nature of the war that resulted, namely an air campaign that disproportionately harmed innocent civilians and their property, rather than the military forces it was seeking to suppress. Despite efforts to avoid directly bombing civilians, the air strikes intentionally targeted civilian infrastructure (so-called dual-use targets) such as bridges, roads, power stations, water purification plants, *etc.*, which greatly impacted the lives and safety of civilians. Moreover, while the warring military factions in Kosovo were not able to conduct major military operations, they were not eliminated or significantly hurt by the air strikes either, and indeed switched to urban fighting tactics and shelling cities from hidden positions which further imperiled civilians. Walzer does not dispute that the underlying cause of NATO involvement was just. He simply argues that NATO should have committed to a ground war which would have saved many innocent civilians from harm, even though it would have increased the risks for the NATO soldiers.

Even if robots did make it easier for a nation to go to war, this in itself does not decide whether that war is just or not. There is, however, a deeper question of political justice lurking here that concerns whether it is desirable to make it practically easier to go to war or not. If we assume that only nations fighting just wars will utilize such technologies, then it would not necessarily be unjust or immoral to develop those technologies. However, history instructs us that all wars involve at least one unjust (or badly mistaken) nation, and so the chance that such technologies will enable future injustices is a real and legitimate concern. Moreover, it is likely that obviously just wars do not need their barriers lowered, and so this function tends to aid the propaganda of aggressors more than that of just defenders. If we agree that the majority of wars are in fact unjust on one side, then any technologies that lower the barriers to entry of war are empirically more likely to start wars period, even if one side has just cause to enter it. Still, this is an argument against militarization in general, and not specifically about autonomous systems and robots, even if they are a dramatic example of it. From an empirical perspective, it is also important to consider why these specific technologies are being heavily invested in rather than other technologies, and if it is primarily due to this propagandistic function, then this should raise concern.

## 3. Autonomous Technology and jus in bello

Walzer claims that just war theory is largely indifferent to the kinds of technology that are used in battle. As far as he is concerned, there are individuals who have a right not to be killed, the innocent civilians, and those who have given up that right by taking up arms, the uniformed combatants. As it is morally permissible to kill uniformed combatants, it does not matter much how one goes about killing them (assuming that one recognizes their right to surrender, *etc.*). However, there are a number of international conventions which do limit the use of specific kinds of weapons, such as chemical, biological and nuclear weapons, as well as landmines, lasers designed to blind soldiers, and other sorts of weapons. There are various reasons for the existence of these treaties, and several principles which determine what kinds of technologies are permissible as weapons of war. In this section, we will consider how autonomous technologies might challenge the standards of *jus in bello*.

Despite Walzer's claims that his theory does not care about the technologies used for killing, he does discuss several specific technologies in terms of how they changed the conventional standards of war. In particular, the use of submarines, aerial bombing and the atomic bomb, all relatively new technologies, changed the accepted conventions of warfare during World War II.

The clearest example of this is the way in which the specific technologies of submarine warfare in World War II served to repeal a centuries-old naval warfare convention. The convention held that there was a moral duty to rescue the surviving crew of a sinking enemy ship once a battle was over. Over the long history of European naval warfare, this convention made sense to all participants, since combat usually occurred in open seas, often hundreds or even thousands of miles from safe harbors, and disabled or sinking ships generally had significant numbers of survivors. From the development of submarines through World War I, this convention held for submarines as it had for all other ships. It was decided during World War II, however, that this convention no longer applied to submarines. The reason for this was that requiring a submarine to surface and conduct rescue operations would make it too vulnerable to

detection from radar and attack from airplanes armed with torpedoes. Additionally, the small submarines (with crews of less than 50) could not spare space for survivors on-board, adequate guards to take officers prisoner, or space to store much rescue equipment (ships sunk by submarines often had more than 1,000 people on board), all of which made rescue efforts challenging and impractical. They could, however, right upset lifeboats and provide food and water, as well as pick up men from the sea and put them in their own lifeboats, but these activities were considered too risky.

While there were some specific and dramatic events that led to the abandonment of this particular convention for submarines, it was largely due to the fact that obedience to the convention was so risky that it would render submarine warfare impractical. The official abandonment of the convention occurred when the German Admiral Doenitz issued the *Laconia* Order in 1942, which expressly directed submarines to not engage in any form of assistance to survivors [1]. Doenitz was tried at Nuremberg for a war crime in issuing this order, but was acquitted of the charge by the judges. The legal decision rested primarily on the fact that because both sides assented to the new convention in practice, the old convention was effectively annulled and submarines no longer had a moral obligation to rescue crews, despite the fact that it was sometimes safe for a submarine to rescue survivors. Walzer believes this is the correct moral interpretation, and accounts for it in his theory as a valid use of the principle of military necessity. That is, it became a military necessity to forgo the convention of rescue in order for submarine warfare to be an effective naval strategy, though this clearly makes naval warfare a more brutal affair.

I believe this deference to military necessity presents a significant weakness in just war theory, as formulated by Walzer, when viewed in the context of technological development. That is, why should we not say that submarines should not be used at all if they cannot be used in a manner which conforms to the conventions of just war? If we cannot argue this way, then there would seem to be a certain kind of impotence to using just war theory to argue against any technology that has the potential to change conventions via military necessity. Not only does this position mean that we have to accept all new technologies and the new conventions that arise from them, but also that we cannot judge the morality of devising various sorts of weapons. The question is: If I can only defend myself with an indiscriminate and disproportionate weapon, because that is the only militarily effective weapon I have, then was I acting unjustly when I chose to arm myself with that weapon rather than another weapon which could be discriminate and proportionate? Do I have a moral duty to invest in weapons that will not tend to put me in positions where I may be motivated to act unjustly (indiscriminately and disproportionately) in the future? If so, could robot soldiers be such a technology?

This failure in Walzer's formulation stems from its loose foundational considerations. While Walzer is willing to assert individual rights as the basis for prohibitions against killing civilians, he mistakenly asserts that soldiers forgo their rights not to be killed by merely taking up arms. Further, he seems to believe that the restrictions on the use of weapons against soldiers, and the rights of soldiers to surrender, rescue, medical aid, *etc.*, are a matter of convention between states that see these conventions as being in their mutual interest. Thus, there is no firm moral foundation to prevent states from abandoning these conventions when they are deemed not to be in their mutual interest. A convention depends only upon both sides assenting to it, as in Walzer's analysis of the Doenitz decision which rests upon the fact that both sides of the conflict observed the same convention.

Apart from convention, Walzer might appeal to moral sentiments in determining the morality of certain military strategies and technologies. To the extent that it derives its principles from moral sentiments, just war theory is an attempt to describe sentiments that occur *after* an event. In the case of many new technologies, we do not really know how that technology will function in the complex socio-technical system of war. Submarines, radar and airplanes armed with torpedoes had never been used together in warfare before WWII, so nobody really knew how they would be used. Indeed, the German admiralty only abandoned the sea rescue convention for submarines in 1942, well into the war. Thus it seems that if we cannot predict military necessity very well, just war theory as it stands cannot tell us much about which technologies might be best left undeveloped.

The more critical issue that just war theory faces is that the conventionalist and sentimentalist interpretations both fail to secure a moral foundation for restrictions on actions against combatants *in bello*. Most generally, if we accept that the combatants on a just side of a war have not waived their rights not to be killed [2], then no conventional agreement between states can waive or trump that right. Similarly, if sailors have a moral right to be rescued after their ship is sunk, then neither the demands of military necessity, nor the technological limitations of submarines, nor the conventional agreements between belligerent navies can waive that right, even if it makes it impractical to respect it. The practicality issue comes into play only when we come to consider the legal restrictions on combat, not its fundamental morality. Thus, we might accept a certain degree of immorality in our laws because of the impracticality of judging and enforcing moral justified laws [2].

The real question then becomes one of what moral rights individuals have against the use of specific technologies, and of the moral duties of states in the development of arms that might be used in hypothetical future wars. Again there is the distinction between fundamental morality and practical law, but it seems possible in principle to develop a moral foundation for arms control and limitations on the design and use of various technologies. While it is well beyond the scope of this essay, it will remain a topic for further research.

### *3.1. Distinguishing civilians & combatants*

On Walzer's interpretation of just war theory, the most fundamental distinction made by just war theory is that between combatants and civilians. While this distinction fails, like military necessity, to find solid moral grounding, it proves quite useful in establishing practical laws for regulating war. This distinction makes it legally permissible, at least sometimes, for combatants to kill enemy combatants. It also makes it possible to say that it is almost never legally justified for combatants to kill innocent civilians. There are, of course, shades of grey. Even combatants retain certain rights, like the right to surrender, and not to be killed unnecessarily. There are also cases in which it is legally permissible to kill civilians—but these cases must meet a very strict and limiting set of conditions, and even those are contentious. Further problems arise in guerrilla and insurgent warfare, in which combatants pose as civilians.

In this section I want to consider several different aspects of the problem of distinguishing civilians and combatants as it relates to autonomous systems. First, the practical ability of autonomous technologies to draw this distinction correctly is crucial. On the one hand, it has been argued that this ability makes the use of such systems *morally required* if they are available. What is more surprising is that it is human rights

groups, such as Human Rights Watch [7], that are making this argument and demanding the use of only "smart" bombs in civilian areas. On the other hand, it is the fear of indiscriminate violence, perhaps mixed with impoverished cultural and social intelligence, that makes robotic soldiers seem particularly dangerous and morally undesirable.

The relevance of the civilian-combatant distinction to robotic soldiers is that if they are to be autonomous in choosing their targets, they will have to be able to reliably distinguish enemy combatants from civilians. It seems that this capability will remain the most difficult theoretical and practical problem facing the development of such robots. While there are technologies for picking out humans based on computer vision, motion and heat patterns, it is extremely difficult to identify particular people, or even types of people, much less to categorize them reliably into groups such as "friend" or "foe," the boundaries of which are often poorly defined and heavily value-laden.

In keeping with the Human Rights Watch argument, there is a line of reasoning which asserts that advanced technologies have the potential to be superior to human capabilities. Arkin [6] argues that if we can achieve the proper discriminatory capabilities in robots, they may very well be *morally superior* to human soldiers. The argument maintains that if machines are better able to discriminate civilians from combatants, then they will make fewer mistakes than humans. Moreover, because it is a machine, it will not feel the psychological and emotional stress of warfare, and thus will not be inclined to commit war crimes or atrocities as humans under such pressure might. Thus, there is not only a moral obligation to use such systems when they are available, but also to build them (insofar as war is taken as an unavoidable feature of human civilization). This all depends, of course, on the actual abilities of the technology, and the abilities of combatants to fool such systems into misidentification.

### *3.2. "Push-button" wars*

Walzer notes that a radical transformation in our understanding of the war convention occurred with the rise of the modern nation-state. Before this, warfare was largely conducted by individuals who freely chose to participate in a given war, and a given battle. With the rise of the modern nation-state came the power to recruit and conscript individuals into standing armies. Because of this, nearly all individual soldiers lost their freedom to choose which wars and which battles they would fight in, even if they had the choice of whether to volunteer for military service. The consequences for moral estimations of conduct in war were thus reshaped by our knowledge that many of the actual combatants in war are not there freely, and thus deserve a certain degree of moral respect from their own commanders as well as from enemy commanders. While it is permissible to kill them, they still have the right to surrender and spend the rest of the war as a prisoner. It is also morally required that commanders seek out ways of winning battles that minimize killing on both sides. That is, the lives of the enemy still have moral weight, even if they weigh less than the civilians and combatants on one's own side. And the lives of one's own soldiers also count in a new way. Whereas it might be moral to lead a group of soldiers on a suicidal charge if they all volunteer for that charge, ordering conscripted soldiers into such a charge is usually immoral. In the same way, it is deemed highly honorable to throw oneself on a grenade to save one's comrades, but not to throw one's comrade onto the grenade—the autonomy of the individual soldier to choose his or her fate has moral implications.

The use of autonomous systems may similarly change our conception of the role of soldiers in war, by fully realizing a "push button" war in which the enemy is killed at a distance, without any immediate risk to oneself. This approach to war could be deemed unjust by traditional conventions of war because those doing the killing are not themselves willing to die. This principle is fundamental because it powerfully influences our sense of fairness in battle, and concerns the nature of war as a social convention for the settling of disputes. Insofar as it can serve this purpose, both sides must essentially agree to settle the dispute through violence and, by the norms of the convention, the violence is to be targeted only at those who have agreed to fight, *i.e.* the combatants. Thus it is immoral to kill civilians, who have not agreed to fight. This convention is only abandoned in a "total war" in which no actions are considered unjust because the stakes of losing are so high. By fighting a war through pressing a button, one does not fully become a combatant because one has not conformed to the norms of war in which both sides agree to risk death in settling the dispute. The limitations of such a conventionalist notion of just war have been noted above, however, and there would seem to be no deeper moral obligation for a just combatant to risk their own lives in defense of their state.

We could imagine a war in which both sides sent only robots to do the fighting. This might be rather like an extremely violent sporting contest in which the robots destroy each other. For this to actually count as a war, and not merely a sport, however, political decisions would have to be made as a result of this competition, such as ceding territory. While it might seem unlikely that a nation would simply give up its territory or autonomy once its robots were destroyed, this is not an unreasonable or impossible outcome. It might also be deemed moral to fight to the last robot, whereas it is generally not deemed moral to fight to the last human. While many nations have surrendered after the crushing defeat of their armies but before the actual conquest of their lands, it would seem likely that a state might continue fighting with humans after its robots have been destroyed, rather than simply capitulate at that point. In general, I think it is fair to say that an exclusively robotic war might even be a highly preferable way of fighting to what now exists. In its most extreme form, we could even imagine a decisive war fought without a single human casualty.

Such a war would not be completely without its costs and risks, however. First, such a war would have to take place somewhere, and it seems likely that the destruction of natural resources and civilian property would be highly likely in most locations. As the most common military objectives are to hold cities and towns, there is both the risk of harming civilians in the course of fighting, and the problems of holding towns, and thus controlling and policing civilian populations with robots. There would also be the cost in terms of the time, money and resources devoted to building up these robot armies.

At the other extreme lies the completely asymmetric "push-button" war. Thanks to science fiction and the Cold War, it is not hard to imagine an autonomous military system in which the commander needs only to specify the military action, and press a button, the rest being taken care of by a vast automated war machine. We could even imagine a civilian government that has completely replaced its military with a fully automated system, perhaps designed and staffed by civilian technicians, but one that did not require any uniformed soldiers to operate. Such a system would, I believe, seriously challenge the conventional concept of war.

In a completely asymmetric war, in which one side offers no legitimate uniformed combatants in battle, but only robots, our moral sentiments could be profoundly upset.

If one nation fights a war in which its soldiers never appear on the battlefield, offering no opportunity for them to be killed, then the combatants are all machines and the humans are all civilians. As in a guerrilla war, one side presents no legitimate human targets to be killed. A legitimate army would not have any opportunity to reciprocally kill the soldiers of their opponents in such a situation (and could only inflict economic damage on their robots). This could thereby be interpreted as a fundamental violation of the war convention itself, like showing up for a duel in armor or sending a proxy, and thereby as a nullification of the associated conventions. Seen another way, such a situation might also be presented as an argument in favor of terrorism against the civilians who sit behind their robotic army. It could be argued that because such an army is the product of a rich and elaborate economy, the members of that economy are the next-best legitimate targets. This possibility should alert us to the unsuitability of conventions and moral sentiments, rather than individual rights, as a basis for just war theory, since we would not want a theory of just war which legitimizes terrorism.

If we instead see the foundations of just war as deriving from individual rights, it would be unreasonable to insist that a nation fighting for a just cause is obliged to let an unjust aggressor kill its citizens even though it has the technological means of preventing this. Indeed, outside of Walzer's interpretation of the moral equality of soldiers, we do not expect a technologically superior nation to refrain from using its available technologies simply because they give too great of an advantage, nor do we expect a larger army to use special restraint in fighting a smaller army out of a moral sense of fairness. Similarly, as long as the robot army is no more likely to cause unjust harms than a human army, it would seem to offer a superior military advantage in limiting the risks to one's own citizens.

There is a compelling rationale for a nation desiring to defend itself without risking human lives. That is, a nation could quite reasonably decide that it does not want its children to be trained as soldiers or sent to war, and so develop a technological solution to the problem of national defense that does not require human soldiers, namely a robot army. In such a case it would not seem to be immoral to develop and use that technology, and we might go even further and say it is morally required for that nation to protect its children from becoming soldiers if it is within their technological capacity to do so. If an aggressor invaded this nation, I do not think many people would raise a moral objection to their using robot soldiers to defend themselves.

Of course, the push-button war is already available in a certain sense, namely for those countries with superior air forces and a willingness to bomb their enemies. The practical consequence of such wars is the asymmetric war in which one side is so obviously technologically powerful that it does not make much sense for the opposition to face it in the traditional manner. The result is often guerrilla warfare, and sometimes terrorism. The advent of robot armies may further exacerbate such situations, but would not seem to be fundamentally different. Their development and employment should, however, take into consideration that these are likely responses to the use of robot armies, even if they are not morally just responses.

## 4. Conclusions

Ultimately, just war theory concludes that the use of autonomous technologies is neither completely morally acceptable, nor is it completely morally unacceptable under Walzer's [1] interpretation of just war theory. In part this is because the technology,

like all military force, could be just or unjust, depending on the situation. This is also, in part, because what is and is not acceptable in war, under this interpretation, is ultimately a *convention*, and while we can extrapolate from existing conventions in an attempt to deal with new technologies, like autonomous killing machines, this process can only be speculative. It is up to the international community to establish a new set of conventions to regulate the use of these technologies, and to embody these in international laws and treaties. Such a process can be informed by Walzer's theory, but his approach is to appeal to conventional practices as the ultimate arbiter of military necessity when it comes to technological choices. In light of this, we may wish to extend or revise the theory of just war to deal more explicitly with the development and use of new military technologies. In particular, we might seek to clarify the moral foundations for technological arms control, perhaps upon individual rights or another solid moral ground. Such a theory might also begin to influence the practical control of autonomous weapons systems through international laws and treaties. I believe that this would be a promising approach for further work.

## References

[1] M. Walzer, Just and Unjust Wars: A Moral Argument with Historical Illustrations, Basic Books, NY, 1977.
[2] J. McMahan, The Sources and Status of Just War Principles, Journal of Military Ethics, 6(2), 91-106, 2007.
[3] J. S. Canning, A Concept of Operations for Armed Autonomous Systems: The difference between "Winning the War" and "Winning the Peace", presentation at the Pentagon, 2007.
[4] K. Capek, Russum's Universal Robots (RUR), 1921.
[5] L. Alexander, The Doomsday Machine: Proportionality, Prevention and Punishment, The Monist, 63, 199-227, 1980.
[6] R. C. Arkin, Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture, Georgia Institute of Technology, Technical Report GIT-GVU-07-11, 2007.
[7] Human Rights Watch, International Humanitarian Law Issues in the Possible US Invasion of Iraq, Lancet, Feb. 20, 2003.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Limits to the Autonomy of Agents

Merel NOORMAN
*Philosophy Department, Faculty of Arts and Social Sciences, Maastricht University, The Netherlands*

**Abstract**: The concept of *autonomous artificial agents* has become a pervasive feature in computing literature. The suggestion that these artificial agents will move increasingly closer to humans in terms of their autonomy has reignited debates about the extent to which computers can or should be considered autonomous moral agents. This article takes a closer look at the concept of autonomy and proposes to conceive of autonomy as a context-dependent notion that is instrumental in understanding, describing and organizing the world. Based on the analysis of two distinct conceptions of autonomy, the argument is made that the limits to the autonomy of artificial agents are multiple and flexible dependent on the conceptual frameworks and social contexts in which the concept acquires meaning. A levelling of humans and technologies in terms of their autonomy is therefore not an inevitable consequence of the development of increasingly intelligent autonomous technologies, but a result of normative choices.

**Keywords.** Autonomy, artificial agents, moral responsibility, self-regulation

## Introduction

To what extent can computers become autonomous entities working and collaborating with humans? This is a central question in debates about the now pervasive notion in Computer Science of *artificial intelligent agents.* In this paper, I consider the notion of artificial agents in a broad sense, including robots, as well as digital entities, such as email assistants or components in software programs [1-3]. Autonomy is the one feature that is commonly identified as defining an agent in its most abstract form. As a key feature, autonomy is also one of the most contested concepts. Despite several attempts in the agent literature to find a proper and all-inclusive definition, autonomy remains an elusive and ambiguous concept, much like the term agent itself [3,4]. Nevertheless, the implicit analogy with human autonomy supports a prevalent rhetoric in agent discourse of future worlds, in which computers will become animate entities that independently set out to accomplish their own goals; as if they have a life of their own. At the same time, this rhetoric presents future artificial agents as *delegates* of or *collaborators* with humans. These agents would go out onto complex information networks and into physical environments to perform tasks "on behalf" of humans. Personal digital assistants will, for instance, manage our daily appointments and our

communication with others [5,6]. Social robots will take care of medical patients and our elderly [7,8]. Military autonomous vehicles will take the place of human soldiers and go out into combat on our behalf [9].

The discussion on the limits to the autonomy of computer technology is a recurrent theme in agent discourse. The prospect of integrating and incorporating autonomous agents in daily practices presents questions about the distribution of responsibility and accountability [10,11]. In particular, the notion of progressively independently-acting computational entities performing complex tasks in ways no longer directly traceable to or comprehensible for human actors raises concerns about the extent to which these technologies can or should be delegated control. Recently, these concerns have become the focus of considerable attention in agent research as well as in Computer Ethics. Various authors have considered the question whether artificial agents can or should be considered autonomous moral agents [12-15]. Underlying this question is the contentious notion that future technologies, and in particular artificial agents, will move closer to humans, as their increased competences challenge and dissolve traditional boundaries between humans and technologies. A number of participants in these debates have endeavored to identify essential qualities or properties of autonomous moral agents, either to formalize and implement these features or to dismiss the possibility of autonomous artificial agents on a par with humans. The aim of this paper is to offer a different perspective on the concerns raised by the idea of autonomous artificial agents.

"Essentialist" discussions run the risk of loosing sight of the interdependencies between humans and technologies [16]. Literature from the field of technology studies and philosophy of technology has drawn attention to the mutual constituting processes between humans and technologies [17-19]. This literature has highlighted the influence of the social and cultural structures in which technology is developed and used. Moreover, it has emphasized the profound effects that technologies have on the way humans act as well as on how they understand the world and what it means to be human. As Katherine Hayles notes: "humans create objects , which in turn help to shape humans"[20].

An important insight that these studies of the interdependencies between humans and technologies have offered is that as part of the processes of mutual shaping our conceptions of humans and technologies are continuously contested and redefined [21]. A preoccupation with an analysis of inherent properties distracts from an informed debate about which, why and how particular boundaries between humans and technologies should be maintained or dissolved. I will therefore not attempt to formulate a definition of autonomy that excludes or includes humans and artificial agents. Instead, I consider autonomy to be a context-dependent malleable concept that is instrumental in understanding, describing and organizing the world.

A closer look at the concept of autonomy in agent research reveals that the tension between the optimistic promises of autonomous artificial agents and the various concerns that they produce arises from the confrontation between two different conceptions of autonomy, rooted in distinct conceptual systems. As I hope to show, this tension indicates that a persisting boundary between humans and technology remains that leaves the human as the ultimate morally responsible party. This is not because of some inherent qualities of either humans or technologies. Rather, the persistence of this boundary originates in the social and cultural structures in which human/technology relationships take shape. Consequently, a change in human/technology relationships that would make the analogy between human autonomy and autonomy of computers

less contested would require a major shift in how we think about technologies as well as in our more fundamental beliefs about moral responsibility.

## 1. Autonomy as Self-regulation

The concept of autonomy appears in a substantial part of agent research as one of a range of abstraction tools that are instrumental in defining agents as a distinct software engineering and development approach. It provides a design metaphor to characterize how one part of a program relates to other parts [22]. Thinking in terms of autonomous agents allows agent researchers and developers to describe, in a high-level abstract sense, a computational entity (e.g. a software component or robot) as a self-contained, self-regulating, interactive unit that operates in some physical or digital environment. The emphasis on the autonomy of agents positions agent-oriented approaches within a spectrum of software engineering methods, such as *Object Oriented Programming, WebServices* or *distributed AI*. Agent researchers Michael Wooldridge and Nick Jennings, for instance, contrast agents as components in a software program to 'objects'[3]. According to Wooldridge and Jennings, objects can be thought of as having some control over their internal state in that this state can only be accessed or modified through the methods that the object provides. An object *x* can 'tell' another object *y* what to do by invoking a method *m* provided by *y*. The object *y* has no control over whether the method is executed or not. In contrasts, agents can only be 'asked' to perform an action. Wooldridge and Jennings assert that agents are thought of as having control over their own actions as well as over their internal state, in the sense that they are not externally directed in the generation and completion of their goals or their decision making processes by other software components. Unlike objects, agents are thought of as 'requesting' other agents to perform an action. The decision to act upon this request is left to the recipient.

The concept of autonomy in agent research also serves to relate artificial agents to humans. A number of research projects on autonomous agents are driven by the explicit objective of taking the human "out of the loop". For tasks that are too complex, too dangerous, or that require accurate time-critical control, increased autonomy of computer technology is an appealing feature. For instance, NASA has a long and diverse track record when it comes to pursuing the development of autonomous robots that can operate at considerable distances from earth with minimal human intervention. It has adopted the term autonomous agents to describe technologies under development that are intended to operate without explicit direct and continuous human intervention, such as the software that operates the *Earth Observing 1* (EO-1) spacecraft [23,24]. The objective in such projects is to minimize the requirement for a human operator in sensing, processing, acting and control loops. The focus is not on analyzing the dynamics of the human/agent relationship; rather it is on isolating the actions performed by technology from this relationship. The kind of autonomy pursued, here, takes the form of relative *closure* of the system's organization, in the sense that physical actions, the decision-making and information processing take place in a *closed control loop* circumscribing only the spacecraft and its immediate environment. The technology-centred view in these projects positions artificial agents and humans as two separate but sporadically interacting systems, where the human is reduced to a peripheral and insignificant element to a largely technological system.

A conception of an autonomous agent as a self-regulating entity is reminiscent of more conventional functional ways of thinking about *automation,* rooted in a cybernetic and information theory tradition. Full autonomy is on the far end of a continuous scale of increasing automation, where automation can be regarded as the use or introduction of machines or computers that are delegated tasks to complete without direct and continuous human control [25]. The level of autonomy is related to the level of control that the machine has over the execution of a process as compared to how much human intervention is required. Higher levels of autonomy are attributed to those automated systems (machines or computers) that are left to perform tasks on their own, and have the *authority* over these processes, i.e. humans have neither the need nor the ability to intervene. Defined as such, autonomy describes an observable and measurable property of a control relationship between an entity and other entities, disconnected from any moral or normative connotations. The notion that an automated technology can act for extended periods of time on its own has no moral implications for the technology itself; it does not attribute to it certain rights or obligations.

In the contexts described above, autonomy is an abstract and formalized concept, instrumental in understanding and developing computer systems in isolation from their connections to humans. This functional conception of autonomy as self-regulation supports an abstract conceptual framework to study particular technical and formal aspects of computer systems. During the process of abstraction, human activity is filtered out. The work performed by researchers and engineers as well as the eventual activities of the humans working with the technologies do not appear in accounts that embrace this conception. In practice, however, technologies do not operate in isolation. They become part of human social organizations and culture, in which a notion of autonomy with moral connotations is prevalent.

## 2. Delegation of Control

The delegation of control in practice places technologies in a continuous dependency relationship with humans, as decisions have to be made with regards to human activity. This is the point where that nagging feeling of loss of control starts to become an issue, as it brings up concerns about the distribution of responsibility and accountability. Scholars critical of techno-enthusiastic discourses about the promises of increasingly autonomous computer technology argue that the idea of autonomous agents encourages the user to attribute a kind of decision-making capacity to the computer that sits uncomfortably with the distribution of responsibility and accountability in daily life [10, 15]. What happens when things go wrong? Can we ask of a pilot to feel responsible for the decision-making of the automated controls of an airplane when she has no direct control or overview of the behaviour of the plane? Moreover, the notion of artificial agents pursuing their own goals independent of human control raises concerns about issues of transparency and trust. How can we *trust* these agents to perform 'on our behalf'? These concerns highlight the tension between autonomy as an abstract functional concept and a more common conception of autonomy in Western contemporary society.

The idea of autonomy is an essential element of liberal democratic traditions as these traditions are predicated on the notion of a person as autonomous agent [26]. As a defining feature of a person, it is a concept that serves to denote that capacity that most people like to think they possess, i.e. the ability to make their *own* decisions based on

their *own* authentic independent motivations. These traditions inherit from Kantian moral philosophy. In a Kantian way of thinking a person is autonomous if he or she acts according to *universal moral rules* or *principles* [27]. Thus, an autonomous person is not directed in her actions by external or internal influences, such as desires or consequences. She acts for moral reasons and because it is an objective obligation to do good. Because an autonomous person is necessarily a *rational* being, she can *rationally* determine which moral principles are authoritative. This conception of autonomy still underlies our contemporary beliefs about what it means to be human and serves various purposes [26]. In liberal democratic traditions a 'normal' person is *assumed* to have the capacity for autonomy, and should not be significantly inhibited in her condition to exercise this capacity. A person has the right and the obligation to act as an autonomous agent [28]. A moral conception of autonomy is expressed in the foundations of our legal system, in human rights, but also in our daily treatment of other people. In our activities, we assume a person makes decisions based on independent thought processes, and excuse or fault them when they fail to do so.

In contemporary Western society autonomy is, thus, more than an element of an analytical description of human beings, as separate from other entities. It can not be considered in isolation from the position of a person in relation to her social, economic, political, juridical and ethical context. In this context, it becomes a non-quantifiable assumed property that serves as an organizing principle and has a prescriptive quality. In contrast, the meaning of the concept of autonomy in the context of automation is primarily a formalized model of an observable relational property of a system that describes a gradual scale of organizational dependence between two entities. The most notable difference between these two meanings of autonomy is that the first has strong normative connotations, whereas in the second the moral dimension of the relationship of control between human and machines plays an inconsequential role. It is this difference that generates a tension, when the concept is applied to artificial agents in the context of human activity.

To resolve this tension, several agent researchers have proposed to extend the competences of artificial agents to enable them to reason about the moral dimensions of their actions [12]. In particular, one suggested strategy to deal with concerns about loss of control, is to move agents closer to humans by formalizing and digitizing those abilities or qualities that enable humans to be part of a social organization; those abilities that allow them to collaborate on the basis of an understanding of shared norms and social rules[1]. For instance, the cognitive scientists Cristiano Castelfranchi argues that if artificial agents are to be embedded in a complex socio-cultural environment they should be capable of understanding the mechanisms of what Castelfranchi calls "social order" in order to effectively support human activity [30]. They have to be able to 'understand' the informal processes in spontaneous and "bottom-up" interpersonal relationships that give rise to social order. He argues that this requires a "formalisation" of these informal dynamic processes, in addition to the formal mechanisms such as rules, regulations, protocols and legislation.

One problem with formalizing the mechanisms of social order or other types of mechanisms is that such approaches tend to exclude existing orderings and fundamental beliefs about humans and technologies from analysis. They conceptually place humans

[1] In agent research normative models are also used as abstraction tools to model communication protocols between artificial agents [29]. In this article, I am particularly concerned with those approaches that formalize moral rules and norms to interact with humans.

and artificial agents on the same level, assuming that if the mechanisms supporting social order and autonomy can be formalized, then humans and agents are treated as equal agents. However, the acquisition of knowledge about moral behaviour, as well as the use of this knowledge, is problematic when it comes to machines. This is not only because fifty years of research into the possibilities of artificial intelligence have demonstrated the extent and difficulty of the objective of building such technologies; it is also because deeply rooted beliefs about the boundaries between humans and technologies in Western cultures favour particular human-technology configurations.

## 3. Persistent Asymmetries

Human Factors researcher Victor Riley asks: "With all the complexities surrounding human interaction with automation, and recognizing that automation can perform many tasks more precisely and reliably than human operators can, one may wonder why we don't just automate the operator out of the process altogether" [31]. The answer to this question, Riley suggests, is that as long as we feel the need to blame someone when things go wrong we will assign a responsible human operator. Riley's comment exemplifies a persistent conceptual and normative asymmetry between humans and technologies. This asymmetry renders humans the ultimate *morally responsible* party. We can hold something 'accountable', by replacing or modifying it, but the search for moral responsibility does not stop there. Chains of responsibilities are traced back to human operators, developers, manager, or even politicians: there is a bug that needs to be fixed, or developers did not have enough training, the impact of technology was not accurately anticipated. The tendency will be to hold those developing, using and managing technologies ultimately responsible and accountable for these failures. This anthropocentric bias has consequences for the delegation of control and actions.

In their analysis of the delegation of control and action, the sociologist Harry Collins and the philosopher Martin Kusch (from now on CK) highlight the asymmetrical treatment of humans and technologies [32]. They observe that humans delegate only a particular kind of action to technology. CK conceive of actions and their meaning as intimately tied to a culture or, as they call it, in reference to Wittgenstein, a *form of life*. In contrast to behaviour an action is more than a reflex. Rather, it is an *intentional operation*[2] composed of a set of behaviours the meaning of which derives from a cultural context constituted by shared concepts and actions.

CK distinguish between two kinds of actions: *mimeomorphic* and *polimorphic* actions. Mimeomorphic actions can be reproduced by an individual outside of a cultural context. This individual can mimic the behaviours that constitute the action without understanding the significance of these behaviours. The exact reproduction of the behaviours appears to reproduce the action to a member of a form of life who understands the meaning of the action. This kind of action can potentially be delegated to machines, they argue. A soccer robot does not have to understand the rules of a

[2] Collins and Kusch let the distinction between action and behaviour coincide with the distinction between natural and social kinds. Both natural and social kinds have a self-referential component. But whereas in natural kinds the reference extends outwards to something that exists independent of the reference, social kinds exist solely by virtue of its reference to itself. The self referential component in a natural kind such as "mountain" consists of the collective agreed criteria that classify something as a mountain. A social kind, such as "money", is "exhausted by the self-reference".

soccer game in order to play it, and neither does a traffic light system have to understand the significance of the behaviours it is performing.

Polimorphic actions, CK maintain, can only be performed by a member of a society or community (*polity*). They are "rule bound" actions, in the sense that there is a collectively constituted right and wrong way to do the action. However, these actions can not "be specified by listing the behaviours in terms of which it could be carried out", as it is the way that behaviours are arranged that makes them meaningful within a society or a community [32]. If the behaviours were to be copied by an outsider it would not be the "same" action, as the same polimorphic action can be performed in many different ways. For example, the mimeomorphic action of riding a bike or driving a car can be described by a set of behaviours. A robot could be made to perform these behaviours[3]. But the polimorphic action of riding a bike through city traffic requires an understanding of the meaning of this action in order to carry it out correctly. A set of rules cannot exhaustively describe the possible events that would enable appropriate action under the contingent circumstances in traffic. You need to understand how to appropriately deviate from the formal rules.

A crucial difference between mimeomorphic actions and polimorphic actions is the extent to which members of a form of life are "indifferent" to the variations in the way behaviours are carried out. It is at the point where we start to care about these variations that an action becomes polimorphic. At this point the meaning of the behaviours becomes negotiable, and it takes a member of a form of life to participate in this negotiation. Mimeomorphic actions unlike polimorphic actions, therefore, can be delegated to a machine, because we do not have to negotiate or build a shared understanding with the machine.

Although, CK are not specifically concerned with morality or moral responsibility, their distinction between polimorphic and mimeomorphic actions illustrates an anthropocentric bias that is prevalent in Western contemporary societies. In these societies, to be held morally responsible for something implies that one is assumed to know the difference between right and wrong, and could have acted differently. We can therefore blame or praise the person. This conception of moral responsibility attributes an ability to humans that distinguishes them from mechanical determined systems: humans can deviate from rigid rules and protocols to perform their actions in an *appropriate* way. In order to act *appropriately* in unpredictable situations, it is not only necessary to have a static model of what actions are appropriate; it also requires the ability to make judgments in contingent situations based on an understanding of cultural knowledge and the social context; an understanding that, according to CK, can only be gained through "socialization"[4]. This could lead to a discussion about the required essential qualities needed to be socialized, such as mental states or intentionality. But what is equally problematic for those that pursue a levelling of humans and artificial agents, is that besides having the ability to be part of this form of life, it is necessary to be *considered* an entity capable of making context-dependent judgments. A cycling robot that manages to navigate successfully through city traffic

[3] Murata Manufactoring in Japan created "Murata Boy": a self-balancing bicycling robot (www.marutaboy.com).

[4] Arthur Kuflik offers a perspective on how socialization is part of moral responsibility. He identifies a particular kind of moral responsibility, he calls "moral accountability responsibility". He argues that we consider individuals morally responsible agents if they can not only give an explanatory account of themselves, but can also engage in a discussion about the appropriateness of their comportment and be willing to acknowledge, apologize and make amends for their possible errors in judgment [11].

can be perceived as understanding the rules of behaviour. Yet, if an accident would occur, this understanding seems to dissolve, as the designers or other human actors that allowed the robot to operate in city traffic will be held morally responsible.

Regardless of whether or not humans or machines are capable of something as mysterious as intentional behaviour, what CK's dichotomy signals is a deeply-rooted belief about what it means to be an autonomous person. Whether or not individuals really have the capacity or are actually given the right, in liberal-inspired Western societies, the concept of autonomy acts as fundamental organizing principle. As an ideal or organizing principle it has an impressive normative weight. It is instrumental in negotiating the boundaries between freedom and determined behaviour, between moral responsibility and causal responsibility, as well as between humans and non-humans. It provides the conditions to blame or praise a person for her actions, because she is considered to be able and in the position to voluntarily decide upon acting. As long as a computer is not considered to be an autonomous entity in this sense and humans are, moral responsibility remains a human affair. Levelling humans and technologies in terms of their autonomy amounts to overcoming this bias, but this requires more than extending the competences of technologies. It entails a change in our conceptions of humans and technologies and how they relate.

## 4. Changing Human/technology Configurations

An anthropocentric conception of moral responsibility favours particular human/technology configurations in which technologies are conceived of and positioned as instruments that do things for or on behalf of humans. Such configurations are accompanied by requirements for controllability and predictability, which constrain (but not necessarily determine) the range of behaviours technologies can perform. In these configurations, the closure of the control loop in automation amounts to a process of isolating and formalizing the behaviours required to perform a particular task, meaningful within a narrowly defined context. Nevertheless, the existence of an anthropocentric bias in configuring humans and technologies does not mean moral responsibility can be unproblematically applied. It also does not imply that technological development cannot lead to changes in future human/technology relationships

Although deeply rooted, the notion of humans as ultimate moral authority is not an immutable law of nature that holds under all conditions. Moral responsibility, like autonomy, is a malleable concept that acquires meaning within particular contexts[5]. The tendency to hold humans morally responsible, rather than machines, does not mean that responsibility is attributed in equal measures to all humans. Not every person is considered to be able or in the position to make moral decisions. For instance, in Taylorists-style organizations the need for independent thought or moral decision-making on lower levels of the hierarchy is reduced as much as possible [32]. The extent to which a person can or should be held morally responsible is negotiable within communities, in which various cultural, economic and political interests help shape the role of humans as well as technologies.

[5] In considering the question whether computers can be held morally responsible, philosophical discussions of computing and moral responsibility have explored various aspects of moral responsibility, such as the different senses of responsibility and the distributed character of moral responsibility [10,11,33].

The interactions and interdependencies between humans and technologies in practice add another level of complexity to the notion of moral responsibility. To say that technologies are generally not attributed moral responsibility, is not to say that technology does not play a role in moral action. Although humans shape the space in which technology performs, technologies in turn set conditions on the range of actions humans can perform, often in ways not anticipated in its design. Philosophers of technology and scholars in the field of technology studies have drawn attention to the role of technologies in shaping actions [18,34]. They have argued that technological artefacts facilitate and enable particular actions, while constraining, discouraging and inhibiting others. Technological artefacts affect the decisions that humans make and how they make them, and thus shape moral actions. A speed bump, for instance, can impose moral behaviour on a human driving a car, by forcing her to slow down and adhere to local traffic norms. It enforces particular desirable moral behaviour, while it limits the possibility for the human driver to act otherwise.

Acknowledging the malleability of the concepts of moral responsibility and autonomy sheds a different light on concerns about the limits to the autonomy of artificial agents, as it raises a different question. Namely, how does the ambition to level the competences of humans and technologies affect our conceptions of both humans and technologies? If future artificial agents are to become the entities that many wish or fear they will be, then this would entail a shift not only in how we think about technologies, but also in our beliefs about to whom or what we should attribute a capacity to autonomously reason about moral and social actions. It would mean a world, in which moral responsibility and autonomy no longer serve as the same organizing principles, as they do in modern liberal societies. This would perhaps be possible in a world where robots are seen as capable of autonomous thought and moral reasoning, just like animals and other animate entities; or maybe in a world where humans are perceived of as similarly limited in their autonomy as compared to conventional technologies and can not be held morally responsible. Such a shift in thinking, however, is not an inevitable consequence of technological progress; rather it constitutes a normative choice about how we want to organize the world.

In light of the increasing complexity of technologies and the associated exceedingly difficult tasks of attributing moral responsibility, various authors within the discourse on artificial agents suggest a reconsideration of our traditionally held beliefs about the roles and the nature of humans and technologies. In the field of Computer Ethics, some have argued to extend the class of Autonomous Moral Agents by eliminating any anthropocentric bias in concepts such as accountability and responsibility [13]. Advocates of artificial agents have claimed that the increasing complexity of computer technologies demands new design and development paradigms that support thinking in terms of manipulating emergent system behaviour, rather than specifying the behaviour of a system at every level [35]. Such strategies would broaden the space in which technologies can operate independently and unpredictably. But it would also change our conception of the extent to which humans can be held responsible for the things they make or how they use them.

## 5. Conclusion

In this paper I have argued that concerns about the loss of control associated with the idea of increasingly autonomous agents are generated from an ambiguous notion of

autonomy. Two distinct conceptions of autonomy appear in the discourse on the possibilities and risks of the development of artificial agents. On the one hand, autonomy is a concept inextricably linked with the notion of a moral and rational person rooted in a liberal democratic tradition. On the other hand, as inherited from the cybernetic roots of Computer Science, autonomy is a measurable and observable property of the relationship between biological and mechanical systems and their environments. Although both conceptualizations figure autonomy as a property of an entity in relation to its environment, the first has a strong normative connotation and is intimately tied to the idea of what it means to be human. In particular, this conception of autonomy is strongly linked to the notion of moral responsibility.

The anthropocentric bias in the notion of moral responsibility sets constraints on the space in which self-regulating technologies can operate, as it generates a requirement for predictability, transparency and controllability. This does not mean that some final limit to the autonomy of agents exists; rather the limits are multiple and flexible dependent on the conceptual frameworks and social contexts in which the concept acquires meaning. The malleability of the notions of autonomy and moral responsibility opens the door to alternative human/technology configurations in which moral responsibility does not serve as the same organizing principle. This, however, is not an inevitable consequence of technological progress, or of the development of self-regulating artificial agents.

The limits to autonomy of agents are the subject of an important debate. Not because advances in technologies are making the formulation of robotic laws more pressing, but because when viewed from a broad sociocultural perspective it provides an arena in which debates about moral responsibility and ethical concerns can be re-evaluated. It readdresses questions about what norms and moral principles we value and why. Formulating a normative model for artificial agents to operate under provides the opportunity for designers, managers, users and philosophers to negotiate and reflect upon what assumptions or values are inscribed in the technology.

## References

[1] Zambonelli, F. and Parunak, H.V.D., Signs of a Revolution in Computer Science and Software Engineering, in Proceedings of the 3rd International Workshop on Engineering Societies in the Agents World. 2003, Springer-Verlag: New York. p. 13–28.
[2] Luck, M., McBurney, P., and Preist, C., A Manifesto for Agent Technology : Towards Next Generation Computing. Autonomous Agents and Multi-Agent Systems, 2004. 9(3): p. 203-252.
[3] Wooldridge, M. and Jennings, N., R., Intelligent Agents: Theory and Practice. The Knowledge Engineering Review, 1995. 10(2): p. 115-152.
[4] Franklin, S. and Graesser, A., Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. Lecture notes in computer science, 1997. 1193: p. 21-36.
[5] Maes, P., Modelling Adaptive Autonomous Agents. Artificial Life, 1994. 1(1-2): p. 135-162.
[6] Lieberman, H. and Selker, T. Agents for the User Interface. online paper [cited; Available from: http://web.media.mit.edu/~lieber/Publications/Publications.html.
[7] Dautenhahn, K., Socially Intelligent Agents : Creating Relationships with Computers and Robots. Multiagent Systems, Artificial Societies, and Simulated Organizations. 2002, Dordrecht: Kluwer Academic Publishers.
[8] Fong, T., Nourbakhsh, I., and Dautenhahn, K., A Survey of Socially Interactive Robots. Robotics and Autonomous Systems, 2003. 42: p. 143–166.
[9] Arking, R. and Moshinka, L., Lethality of Autonomous Robots: An Ethical Stance, in ICRA'07 IEEE International Conference on Robotics and Automation. 2007: Roma, Italy
[10] Nissenbaum, H., Computing and Accountability. Communications of the Association for Computing Machinery, 1994. 37(1): p. 72-80 (9).

[11] Kuflik, A., Computers in Control: Rational Transfer of Authority or Irresponsible Abdication of Authority? Ethics and Information Technology, 1999. 1: p. 173-184.
[12] Allan, C., Varner, G., and Zinser, J., Prolegomena to Any Future Artificial Moral Agent. Journal of Experimental and Theoretical Artificial Intelligence, 2000. 12: p. 251-261.
[13] Floridi, L. and Sanders, J., On the Morality of Artificial Agents. Minds and Machines, 2004. 14(3): p. 349-379.
[14] Stahl, B.C., Information, Ethics, and Computers: The Problem of Autonomous Moral Agents Minds and Machines, 2004. 14: p. 67-83.
[15] Johnson, D.G., Computer Systems: Moral Entities but Not Moral Agents. Ethics and Information Technology, 2006. 8: p. 195-204.
[16] Sack, W., Artificial Human Nature. Design Issues, 1997. 13: p. 55-64.
[17] Bijker, W.E., Hughes, T.P., and Pinch, T., The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. 1987, London: MIT Press.
[18] Latour, B., Where Are the Missing Masses? The Sociology of a Few Mundane Artefacts, in Shaping Technology/Building Society: Studies in Socio-Technical Change, Bijker, W. and Law, J., Editors. 1992, The MIT press: Cambridge Massachusetss. p. 225-258.
[19] Verbeek, P.P., De Daadkracht Der Dingen : Over Techniek, Filosofie En Vormgeving. 2000, Amsterdam, The Netherlands: Boom.
[20] Hayles, N.K., Computing the Human. Theory, Culture & Society : Explorations in Critical Social Science, 2005. 22(1): p. 132.
[21] Haraway, D., Simians, Cyborgs, and Women. 1991, New York: Routledge.
[22] Luck, M., et al., Agent Technology Roadmap: A Roadmap for Agent Based Computing. 2005, the European Coordination Action for Agent-Based Computing.
[23] Chien, S., et al. Lessons Learned from Autonomous Sciencecraft Experiment. in Autonomous Agents and Multi-Agent Systems Conference (2005). 2005. Utrecht, Netherlands.
[24] NASA. Autonomous Sciencecraft Experiment. 2006 [cited 2006 January 31]; Available from: http://ase.jpl.nasa.gov/.
[25] Sheridan, T.B., Telerobotics, Automation, and Human Supervisory Control. 1992, Cambridge, MA: MIT Press.
[26] Christman, J., Autonomy in Moral and Political Philosophy, in The Stanford Encyclopedia of Philosophy, Zalta, E.N., Editor. 2003.
[27] Hill, T., The Kantian Conception of Autonomy, in The Inner Citadel : Essays on Individual Autonomy, Christman, J., Editor. 1989, Oxford University Press: New York, N.Y. p. 91-105.
[28] Feinberg, J., Autonomy, in The Inner Citadel, Christman, J., Editor. 1989, Oxford University Press. p. p. 27-53.
[29] Elio, R. and Petrinjak, A., Normative Communication Models for Agent. Autonomous Agents and Multi-Agent Systems, 2005. 11(3): p. 273-305.
[30] Castelfranchi, C., Formalising the Informal: Dynamic Social Order, Bottom-up Social Control, and Spontaneous Normative Relations. Journal of Applied Logic, 2003. 1(1-2): p. 47-92.
[31] Riley, V., What Avionics Engineers Should Know About Pilots and Automation. Aerospace and Electronic Systems Magazine, IEEE, 1996. 11(5): p. 3-8.
[32] Collins, H. and Kusch, M., The Shape of Actions: What Humans and Machines Can Do. 1998, Cambridge, Massachusetts: The MIT Press.
[33] Coleman, K.G., Computing and Moral Responsibility, in The Stanford Encyclopedia of Philosophy (Spring Edition 2005), Zalta, E.N., Editor. 2005.
[34] Verbeek, P.P., Materializing Morality. Science, Technology and Human Values, 2006. 31(3): p. 361-380.
[35] Zambonelli, F. and Parunak, V., Towards a Paradigm Change in Computer Science and Software Engineering: A Synthesis. The Knowledge Engineering Review, 2003. 18(4): p. 329-342.

This page intentionally left blank

# Part III

# Mind and World:
# Knowing, Thinking, and Representing

This page intentionally left blank

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Formalising the 'No Information without Data-representation' Principle

Patrick ALLO
*Centre for Logic and Philosophy of Science, Vrije Universiteit Brussel, Belgium*
*IEG, Oxford University, UK*
*GPI, University of Hertfordshire, UK*

**Abstract.** One of the basic principles of the general definition of information is its rejection of dataless information. In general, it is implied that "there can be no information without physical implementation" [1]. Though this is usually considered a commonsensical assumption, many questions arise with regard to its general application. In this paper, a combined logic for data and information is elaborated, and specifically used to investigate the consequences of restricted and unrestricted data-implementation principles.

**Keywords.** Data, semantic information, epistemic logic

## 1. Epistemic Logic and an 'Information First' Approach to Epistemology

When we theorise about the relation between data and information, we automatically engage a debate whose ramifications are not confined to a single scientific or philosophical domain. This surely does not facilitate such an enterprise. For starters, in many domains a tight distinction between these obviously related but separate notions is rarely observed, and in practice the terms of data and information can be used interchangeably. By contrast, the distinction between data and information becomes much more urgent when looked at from the perspective of the physics of computation; especially when, as we find in Landauer, pieces of data are considered as a physical implementation of the information we compute with or reason about. Reverting to looser characterisations, pieces of data tend to be understood as syntactical entities, information rather as a semantical entity. Finally, the distinction between data and information is also crucial to the so-called *data-information-knowledge* hierarchy (henceforth, DIK), one of the central metaphors of information-science and knowledge management [2]. In view of this, the core aim of this paper is to give a sufficiently strong, yet not unrealistic interpretation and formalisation of the relation between data and information, and more precisely between the states of being informed and holding data.

Arguably, it is not a good idea to try to capture the relation between data and information on the basis of the wide-ranging uses of both terms. If feasible at all (which is rather doubtful since in most uses of these terms confusion is more common than an actual insight in the subject) this could only yield a very general theoretical basis. It is therefore preferable to properly constrain the issue by (i) providing general though sufficiently precise definitions of data and information, (ii) fixing the broader

context in which we envisage to use the notions of data and information, and (iii) choosing a suitable formal framework for the formulation of our theory.

As for the characterisation of both basic notions, the general definition of semantic information as *well-formed and meaningful data* is the most obvious starting point [1]. Although it is perhaps still too vague to be used to derive any substantial property of data and/or information, it already settles their distinctness. To a first approximation, any string of symbols should count as a piece of data, but it takes more to qualify as a genuine piece of information. Besides the demand that data need to be well-formed as well as meaningful to count as information, it is also commonly assumed that it is possible to turn data into information. Calling the process of turning data into knowledge the core aim of information-science is itself fairly uncontroversial. Yet, while the development of tools and techniques for doing so is valuable, for present purposes this is not as important as the insight that the adoption of a certain level of abstraction is crucial to obtain well-formed and meaningful pieces of data.

One of the most common methods we use to turn data into meaningful content is the use of a properly functioning language; in more formal terms: a language with a syntax and a semantics. Referring to a level of abstraction, then, is no more than a way of talking about the distinctions that can be effectively expressed in that language. In the case of natural languages such levels of abstraction are fairly hard to determine, but when we turn to formal languages this problem is easily dealt with.

While the received view does not include an explicit reference to the levels of abstraction that serve as an interface between data and information, it is also not opposed to it. Nevertheless, the above description is as far as the consensus about the relation between data and information goes. As soon as we start wondering whether any properly structured piece of data counts as information the opinions promptly diverge. One way to avoid the most common quibbles is to explicitly focus on *declarative, objective and semantic* information (henceforth, DOS). Intuitively, this is meaningful data that can be used to answer a question (data can be exploited by a system as input of adequate queries, see [1]) and is therefore cognitively valuable. As a more precise rendition of the general definition, this is sufficient to satisfy the needs of information-scientists and epistemologists alike. At this point an important question remains unanswered: is the process of turning data into meaningful and declarative content reliable in the sense that it should always yield genuine information?

The standard definition of semantic information is commonly thought to imply that form and meaning alone suffice. Still, two objections to this claim are of particular interest. A first widespread objection suggests that *relevance* is of utmost importance; a second, still controversial objection points to the need for information to be truthful. In each case we should wonder whether 'irrelevant information' and 'false information' are kinds of information rather than no information at all. The question of relevance can quite easily be dismissed. Even if relevance is indeed a property of information we ought to value, we still obtain a more general and conceptually sound theory by considering relevance as a property we do not need to capture the conceptual nature of information itself. By contrast, and despite the debate surrounding this claim [3-6], it is not that obvious to make similar claims about false information. A pragmatic rather than a principled motivation for understanding declarative, objective and semantic information as *truthful, well-formed meaningful data* derives from the constitutive role of acquiring information as a means to attain knowledge. Provided knowledge gets its usual factive reading instead of the ultra-loose sense it gets in information-science, it surely makes sense to apply the same veridical standard to information. In short, while

false information might very well be presented as if it were a solid basis for knowledge, it can never be the stepping stone to knowledge that epistemological theorizing requires.

The above remark is especially relevant for epistemological theories in which the need for information is explicitly affirmed. This includes Drestke's knowledge as information-based belief [7], but it is also true for what I would call an *information first* approach to epistemology. Loosely speaking the latter could be thought of as taking the best of traditional epistemology and the previously mentioned DIK-hierarchy, but this is still a highly misleading description. Formulated as a broader programme, an information-based (formal) epistemology is more conveniently characterised as follows: where $\Gamma$ is an epistemological theory, we say that $\Gamma$ formulates an information-first approach iff (i) it takes the relation of *being informed* as one of the central cognitive attitudes; (ii) it attempts to explain knowledge as a way to be informed rather than as a peculiar way to believe; and (iii) it acknowledges the fundamental connection between being informed and being able to inform as a more basic connection between epistemic states and action than the necessary relation between knowledge and proper assertion.

Moving down the hierarchy, making the relation between data and information precise has a place within this broader programme. In the next section a basic logic for 'being informed' is introduced, and subsequently used to express both the basic features of an information-based epistemology and the problems we face when we include an explicit reference to data in this framework.

## 2. Logics for 'Being Informed' (and some of their rivals)

The formal language used to express the properties of being informed is derived from the standard approaches in modal epistemic logic. It is based on the language of propositional logic augmented with a set of modal operators. In addition to the standard connectives for conjunction ($\&$), disjunction ($\vee$), negation ($\neg$), and implication ($\rightarrow$), we need the following operators: $I_a A$ to express that $a$ is informed that $A$ and $D_a A$ to express that a holds data for $A$ as well as the standard operators for knowledge and belief $B_a A$ to express that a believes that $A$ and $K_a A$ to express that a knows that $A$. Finally, for the purpose of expressing logical relations we use $\Gamma \Rightarrow A$ to express that $A$ is a logical consequence of $\Gamma$ and $B \Leftrightarrow A$ to express that $A$ is logically equivalent to $B$. Using that language, we can elegantly express the basic properties of a purely information-based approach to knowledge, and also focus on the role of data. As a starting point, a fairly minimal characterisation is all we need. The following general properties are proposed:

**non-doxasticism:** knowledge is not defined as a kind of belief, and being informed does not have a belief-component. Hence we reject both the possibility of defining knowledge as $K_a p := B_a p \,\&\, p \,\&\, \ldots$as well as the inclusion of the $I_a p \rightarrow B_a p$ axiom. Crucially, this version of non-doxasticism does not deny that belief is necessary for knowledge, only that it cannot figure as a conjunct in a sufficient definition [8]. As a consequence, the $K_a p \rightarrow B_a p$ axiom which traditionally relates knowledge to belief might figure in a combined logic of knowledge and belief and still be consistent with the 'information first' approach we try to sketch.

**necessary data-component:** it follows from its standard definition that information is a specific kind of data which therefore cannot exist without a data representation. Thus formulated there is nothing objectionable about the fact that we treat information as a kind of data, and that in order to be informed the possession of certain data is required. The problems that nevertheless arise from this principle are of two distinct kinds: one concerns the nature of the data themselves, the other concerns the amount of data required for being informed. Of these, the first issue should not surprise us, for it is one of the standard controversies in the physics of computation [9]. For now, we can easily sidestep the problems introduced by the requirement that information be backed up by a physical data-implementation as long as we ensure that the system we present does not substantially rely on any specific feature of the ontological status of data. The second problem is more intricate, for it only surfaces in connection to the relation of 'being informed'. Despite the fact that our treatment of information as a kind of data intuitively suggests that states of information somehow be supported by data, the explicitation of that support-relation is not really straightforward. Getting the relation between being informed and holding data is precisely the problem that needs to be tackled in the present paper.

**primeness of being informed:** holding a piece of well-formed meaningful data which incidentally happens to be true is not sufficient for being informed. This we see by considering Gettier-like cases in which true consequences can be derived from a set of (partially) false data. As a consequence, it is convenient to treat the state of being informed as a prime state, i.e. a state which satisfies a condition that cannot be decomposed into more basic (but still generally applicable) conditions [8]. Crucially, this also entails that, being informed cannot straightforwardly be identified with holding a piece of information, and that there is no exact match between information as a state and information as a commodity.

As for standard modal epistemic logics [10,11], the systems required to reason about data and information are obtained by adding to standard propositional logic axioms which specify the properties of the modal operators required to express the relations of being informed and holding data. As a logic for being informed, we adopt the proposal from [12] which takes $I$ to be a **KTB**-modality satisfying all but the last of the axioms listed below:

(Nec) If $\Rightarrow A$ then $\Rightarrow I_a A$

(K) $I_a (A \rightarrow B) \rightarrow I_a A \rightarrow I_a B$

(T) $I_a A \rightarrow A$

(B) $A \rightarrow I_a \neg I_a A$

(4) $I_a A \rightarrow I_a I_a A$

Since this is not the place to repeat the reasons for using the resulting system as a proper formalisation for being informed, we should only remark that its main difference with standard epistemic logics is reflected in its exclusion of introspective principles.

Once we have a basic system for $I$, we can start to spell out how it relates to other notions. This is for instance what happens when the so-called entailment-thesis $K_a A \rightarrow B_a A$ is used as the primary principle for connecting the logics of knowledge and belief. Yet, while in such cases two existing logics are adopted from the beginning, this need not be so; we can equally well start from a logic for being informed, and derive the properties of data once we have settled on the proper connecting principles. Inspired by the standard principles used to connect knowledge to belief, three bridge axioms could in principle be considered:

(ID1) $I_a A \rightarrow D_a A$

(ID2) $I_a A \rightarrow I_a D_a A$

(ID3) $I_a A \rightarrow D_a I_a A$

As the state of being informed is characterised by a prime condition, such principles are the only type we need to consider – the relation of holding data is solely investigated as a necessary condition. To begin with, only the first of these is explicitly considered, the latter two are taken into account in due time.

So far the approach sketched above does not immediately yield a logic for data itself, all it implies is that the data one is actually informed of are indeed closed under logical consequence. Nothing is said on the properties of merely holding data. As long as we think about data in the same way as belief, this *logical inertia* of data should presumably count as a flaw for our theory. Intuitively, it surely makes sense to have an independent logic for data, for even if we ought only to reason on the basis of our information, it is often only feasible to reason on the basis of the data one holds. This intuition can only be respected if a rather strong interpretation of the relation of holding data is singled out. The fact is that a sufficiently strong interpretation is not necessarily the most obvious one.

The dilemma we face can be explained by distinguishing between (i) holding mere data, (ii) holding well-formed and meaningful data, and (iii) being informed. Clearly, what we ought to reason about falls under the third category, while what we mostly reason about belongs to the second. If, as often happens, the two latter categories are identified, one could easily conclude that simply holding data has no substantive logical property. Since that option is incompatible with the notion of DOS-information we privilege, another solution must be sought. By being attentive to a common ambiguity in the way we refer to data and information, a more robust solution can be discerned.

## 3. States and Commodities

While it goes without saying that, as a theory, the DIK-hierarchy is hardly helpful; it remains a valuable source of common intuitions. In general it makes sense to define our attitude towards the practice of information-science and knowledge management in terms of an inheritance of the problems it identifies, rather than in terms of the solutions it proposes. In that perspective, two concerns should draw our attention: the

status of the hierarchy itself, and the gradual shift from properties of things to properties of people.

The first of these concerns relates to different ways in which the hierarchy can be read: top-down by identifying necessary relationships like "no knowledge without information" or as "no information without data-implementation"; or bottom-up either by identifying what, say, information lacks to count as knowledge, or by describing the procedures required to "turn data into knowledge." Interestingly, as is the case for epistemologists, the identification of necessary relations (top-down) is less tedious than the formulation of sufficient conditions (bottom-up). However, as the notions of knowledge and information presumed by the hierarchy do not correspond with ours, we should not extensively rely on this analogy. A more interesting concern centers on the tension between treating information as a commodity, a thing that can be stored, passed on, sold etc., and information as something that qualifies the cognitive state of an agent. In the literature, this tension arises in at least two ways. In the view that the notions higher up in the hierarchy define properties of humans whereas those lower in the hierarchy define properties of things, and in the presumably more problematic differentiation between implicit and explicit knowledge.

> The distinction between explicit knowledge and information is even less defensible. If knowledge is a property of people, and embodies prior understanding, experience and learning [45, pp. 9--10], it is difficult to argue that explicit knowledge, recorded in documents and information systems, is any more or less than information. [13]

Yet, while the implicit/explicit distinction is rightly criticised by Rowley, it is therefore not implausible that for each level in the hierarchy we sometimes need to refer to properties of things and sometimes to properties of cognitive agents. For that purpose we might – instead of using the confusing implicit/explicit terminology – want to introduce two distinct levels at which data, information and knowledge are assessed: as states and as commodities.

If we want to reformulate the claim that knowledge is primarily a property of people in the standard philosophical terminology, we should probably say that knowledge is a state of mind. More precisely, when we focus on "knowing that" knowledge can be characterised as a propositional attitude: a relation of subjects to propositions. The contrast with our understanding of data as a thing is obvious, for its reformulation into the philosophical vocabulary does not need to refer to minds, subjects or even propositions. By treating data as constraining affordances representable as strings of symbols, the relation of holding data can conveniently be modelled as a relation between an agent and a particular syntactical object. As such it should probably not be understood as a state of mind, and not be modelled as a relation to a proposition. Finally, if we recall the tension between the relation of being informed as it is modelled by the modal operator $I$ and the relation of merely holding a piece of information, we might conclude that being informed can be conceived as a relation to a proposition *and* as a relation to a particular syntactical object.

The best method for illustrating the consequences of modelling data and information in two distinct ways exploits the contrast between explicit syntactical models and mainstream possible worlds models known from epistemic logic [14]. Roughly, if being informed is modelled on the mainstream approach we have that

$$I_a A \leftrightarrow I_a B \text{ iff } A \Leftrightarrow B$$

That is, if being informed is modelled as a relation towards propositions, for every sentence $A$ out of a set of logically equivalent ones, any $I_a\ A$ expresses one and the same relation. Nevertheless, this should not be confused with claiming that holding a piece of information $A$ is really the same as holding a logically equivalent piece of information $B$. For by making such claims, we automatically revert to a way of reasoning that treats information as a particular commodity. This, in its turn, calls for a different, more refined, model where, given a new relation $\mathcal{I}$, $\mathcal{I}_a\ A$ and $\mathcal{I}_a\ B$ only express the same relation iff $A$ and $B$ consist of one and the 'same' string of symbols. On the face of it, each of these approaches partially captures our intuitive understanding of being informed; while the former largely agrees with the standard approach for knowledge, the latter acknowledges the insight that being informed requires us to hold a particular piece of information. Pieces of information being a kind of data, this way of reasoning about being informed presumably inherits most constraints we pose on our reasoning about data.

If, as suggested above, we leave room for theories and logics of 'being informed' that treat it as a relation to propositions, we face a problem when we try to simultaneously retain the strictly syntactical reading of 'holding data'. Yet, since for data the syntactical reading is by far the most plausible one, it is well worth trying to save it. The problem arises roughly in the following way. Let $I_a$ be a **KTB**-modality, and assume that the relation between being informed and holding data is fully determined by

$$I_a\ A \rightarrow Ð_a\ A$$

Assume now that we try to enforce an otherwise unobjectionable syntactical reading of holding data by, for instance, specifying that $Ð_a\ A$ is only closed under a highly limited set of syntactical manipulations. Thus we get $Ð_a\ [A;B]$ iff $Ð_a\ A$ and $Ð_a\ B$ as a minimal *logical* constraint on holding data (where square braces and semicolons are used to represent the complex syntactical objects obtained by concatenating simpler syntactical objects). Yet, merely constraining $Ð_a$ is not sufficient to ensure the failure of an argument of this form.

(1) $A \Leftrightarrow B$
(2) $I_a\ A$
(3) $I_a\ A \leftrightarrow I_a\ B$
(4) $I_a\ B$
$\therefore\ Ð_a\ A\ \&\ Ð_a\ B$

Viz. if one holds data in virtue of being informed, one's data is thereby also logically closed and $Ð_a$ does not after all succeed to capture a reading of holding data that is systematically weaker than the propositional one. The culprit is nevertheless easily identified, for by formalising the connection between information and data as a straightforward implication, we effectively align them on to the same (propositional) reading. This fact calls for a refinement of the basic data-entailment thesis. An important consideration, in that respect, is the correct interpretation of the view that there is no information without data. That is, if one is informed that $A$ this only means there ought to be some piece of data in virtue of which one is so informed, *not* that there must be a piece of data that somehow perfectly matches one's information that $A$. As a first refinement, the following revised principle is proposed:

**(ID*)** $I_a A$ only if there is an $A'$ where $I_a A' \Leftrightarrow I_a A$ such that $Ð_a A'$

where, in virtue of the failure of the above argument, $Ð_a$ effectively captures a non-propositional reading of holding data. Even then a different but equally objectionable argument is not yet avoidable:

(1) $A \Rightarrow B$
(2) $I_a A$
(3) $I_a A \rightarrow I_a B$
(4) $I_a B$
$\therefore Ð_a A' \,\&\, Ð_a B'$

for some $A'$ and $B'$ such that $I_a A' \Leftrightarrow I_a A$ and $I_a B' \Leftrightarrow I_a B$

As before, since the move from $A$ to $B$ need not be warranted by the weak syntactical manipulations we allow the relation of holding data to be closed under (let $A$ for instance stand for $(\neg p \vee q)$ & $p$, and $B$ for $q$) $Ð_a B'$ is itself not directly implied by $Ð_a A'$ but merely inherited from the closure conditions at the propositional level of being informed. Hence, a further refinement of our connecting principle must be sought.

**(ID**)** $I_a A$ only if there is an $A'$ where $I_a A' \Rightarrow I_a A$ such that $Ð_a A'$

Using this revised principle, the strictly syntactical reading of holding data is duly ensured. For the purposes of the arguments that shall be presented further on, one can uphold that this version basically gets the relation between data and information as well as the syntactical nature of data right. The final refinement we still might want to propose is therefore of less immediate importance, for all it does is integrating the syntactical operations on data within the connecting principle itself.

**(ID***)** $I_a A$ only if there is a complex string of data [$A'$, ...] such that (i) $A'$ can be extracted from [$A'$, ...] ;(ii) $I_a A' \Rightarrow I_a A$; and (iii) $Ð_a$ [$A'$, ...]

Generally speaking, the adoption of this final version has the benefit that no specific assumptions need to be made with regard to the syntactical manipulations that are part of the closure conditions of holding data. This, in its turn, has the advantage that – except for its syntactical representation – the nature of data, and especially its being well-formed, is not fixed by the underlying logic for $Ð_a$. Finally, the separation of mere syntactical manipulations and logical relations results in an agnostic attitude towards the ultimate division of labour between mere symbol manipulation and semantically sensitive inferences that play a role in the process of turning data into information.

Despite the complication the latter two revisions introduce, they are crucial for our ability to reason simultaneously about data as concrete or particular entities, and about information as more general cognitive states. The two levels that are introduced certainly do not rest on *ad hoc* mechanisms. Instead, they serve to make the levels of abstraction at which data and information are evaluated explicit. This results in a more precise reading of the intuitively valid conditions that (i) information needs to be supported by some amount of data, and (ii) being informed involves having precisely that data at one's disposal. This way of limiting the amount of data that is required is particularly valuable in contexts where, because data consist of (or are stored as)

physical entities, dealing with the problem of logical and deductive omniscience becomes even more pressing.

Before we start using these new principles for the design of a combined system of data and information, a few general considerations regarding the method of abstraction should be included. For our purpose, a level of abstraction is best considered as a level of (logical) discrimination; that is, a specified way of, respectively, telling cognitive states (in a formal model) or formulae (in a formal language) apart. Evidently, any account of logical equivalence provides such a level, and the propositional and syntactical accounts of being informed and holding data are obvious examples of this. Besides making the levels of abstraction precise, we also need tools for drawing conclusions at one level of abstraction on the sole basis of our knowledge of what happens at a different level of abstraction. In a simplified formulation, we only have this ability to make inferences that go from one level to another in virtue of a certain degree of inter-LoA coherence. In the present context, being able to do so facilitates the hitherto avoided modelling of data at the propositional level without the loss of the connection with the more fine-grained syntactical approach on which most of our intuitions depend. Its relation to the syntactical approach follows immediately from the above description of the relation between data and information.

**(D**)** $D_a A$ only if there is an $A'$ where $D_a A' \Rightarrow D_a A$ such that $Đ_a A'$

Consequently, while this reintroduces the simple relation between data and information

$$I_a A \rightarrow D_a A$$

it also raises the question of the correct interpretation of this new, less discriminating way of assessing data. A quick and rough answer is this: $D_a A$ expresses the semantic relation of holding data which carries the content that $A$ as opposed to the syntactic relation of holding a piece of data $A$. As such, it does not necessarily refer to a particular object, but only to a state that is warranted by a particular object. The functioning of this notion which can serve as an interface between being informed and holding bare data is best understood in line with the following two principles:

$\triangle$ If every state which satisfies $Đ_a A$ also satisfies $Đ_a B$, then every state which satisfies $D_a A$ also satisfies $D_a B$.

$\triangledown$ If some state which satisfies $D_a A$ does not satisfy $D_a B$, then some state which satisfies $Đ_a A$ does not satisfy $Đ_a B$.

In particular, this entails that if at the syntactical level two sets of data are the same, they remain the same on the propositional level; inversely, if they are distinct at the propositional level, they stay so at the syntactical level. Using only these two principles, several problems can be solved in a more satisfactory manner.

## 4. The Trouble with Meta-data

A mostly convenient context to exploit the method we sketched above relates to the problems posed by meta-data. Basically, meta-data play a role that is similar to that of reflective states in standard epistemic logic. When $B$ expresses the same proposition as $D_a A$, then $B$ is meta-data for $A$: meta-data about the data one holds. Trivially, $D_a B$ then expresses that one holds such meta-data; in short: $D_a D_a A$. With this in mind, one should then ask how $D_a D_a A$ relates to $D_a A$. This relation splits up in two fragments. First we should inspect the prospects and consequences of the principle stating (at the propositional level) that there is no meta-data without data.

$$D_a D_a A \rightarrow D_a A$$

Call the principle in question the 'non-corrupted meta-data' thesis, and note that it is harder to reject than either $D_a A \rightarrow A$ or $D_a (D_a A \rightarrow A)$. The converse of these principles states that one cannot hold data without also having meta-data for it

$$D_a A \rightarrow D_a D_a A$$

As before, we cannot conclude the falsity of this principle, from the sheer fact that we do not hold data for all truths $A \rightarrow D_a A$. Rather, the question becomes whether $a$ should be able to obtain the meta-data for all the data it holds by purely logical means. Call the principle which supports this the 'free meta-data' thesis.

In view of the $\triangle$- and $\nabla$-principles presented above, it should be obvious that the distinctness of $D_a A$ and $D_a D_a A$ cannot be argued for by merely referring to the actual distinctness of states wherein one holds data, and states wherein one holds meta-data. That is, the presumed invalidity of $Đ_a A \rightarrow Đ_a Đ_a A$ and $Đ_a Đ_a A \rightarrow Đ_a A$ cannot be used to infer the falsity of the corresponding principles we are actually interested in. Yet, using the refined data-principles we can approach the problem more carefully along the following lines.

$D_a A$ only if there is an $A'$ where $D_a A' \Rightarrow D_a A$ such that $Đ_a A'$

$D_a D_a A$ only if there is an $A''$ where $D_a A'' \Rightarrow D_a D_a A$ such that $Đ_a A''$

Quite naturally, for holding data to imply holding meta-data, or conversely, for holding meta-data to imply holding data, the above instances of our general principle at least require that they be supported by a single syntactical entity or datum. This at least requires that $A'$ and $A''$ could be the same datum. Two separate cases arise: either that datum can be represented by a purely factual (i.e. propositional) formula, or it can be represented by a data (i.e. modal) formula. To check the former option, we replace $A'$ and $A''$ by the propositional formula $B$; to check the latter, we replace $A'$ and $A''$ by the modal formula $D_a B$. Since $A'$ can easily be substituted for a propositional formula, and $A''$ for a modal one, the following reasoning should work out. If the replacement by $B$ works out, the "free meta-data" principle turns out to be unproblematic; if the replacement by $D_a B$ works out, the "non-corrupted meta-data" thesis is unproblematic. Finally, this requires that for all $A$ there exists some propositional formula $B$ such that, respectively, $D_a B \Rightarrow D_a D_a A$, and $D_a D_a B \Rightarrow D_a A$. Since this approximately brings us

back where we started, this means that given the relation between data-particulars and propositional data we defined, natural constraints on particulars do not fully determine the constraints on propositional data.

This conclusion should, however, not pose a problem for the present enterprise. Rather, it shows that while modeling data, information, and their relation at the propositional level, we enjoy a certain freedom that goes beyond the mere abstraction from actual syntactical specificities. We can also abstract from the difference between data and meta-data, and this is just one of the several choices we face when constructing a formal model. This apparent freedom does not deny that many external considerations can serve as a guide. One such kind of consideration derives from our interpretation of $D_a$. For instance, when $D_a$ is a relation to propositions, we suggest it should be read as "holding data *for* ..." instead of "holding a datum ..." In the same vein, if iterations of $D_a$ intendedly refer to different levels of meta-data, it is somehow incoherent for both $D_a A \rightarrow D_a D_a A$ and $D_a D_a A \rightarrow D_a A$ to be valid. This is especially true if one thinks that even the very basic kind of meta-data an iteration of $D_a$ refers to has a value that exceeds what is already available at the primary level. For if that is the case, then at least $D_a A \rightarrow D_a D_a A$ has to go. By contrast, if one thinks that when the particular datum which initially supported $D_a A$ is destroyed, the propositional content of $D_a A$ cannot infallibly be recovered from that of $D_a D_a A$, then $D_a D_a A \rightarrow D_a A$ should perhaps also go.

With the minimal relation between 'being informed that' and 'holding data for' expressed by **DD**** and a plea for a principled distinction between data and meta-data, a sufficiently large set of constraints for a combined logic for data and information has been obtained. Thus, a system can be fully described on this basis, and the value of some additional connecting principles can also be assessed. Starting with $I_a$ as a **KTB**-modality, we add $I_a \mathrm{A} \rightarrow D_a \mathrm{A}$ as a by now well-motivated expression of the necessary relation between being informed and holding data. Finally, we also specify that meta-data should not come for free, hence $D_a A \rightarrow D_a D_a A$ is not allowed to come out as valid. This is the main guide for the evaluation of the acceptability of ID2 and ID3 – the two remaining connecting principles whose status we did not yet settle. As a matter of fact, they can both be dismissed in an entirely uncontroversial way. Namely, since $I_a D_a \mathrm{A} \rightarrow D_a D_a \mathrm{A}$ and $D_a I_a \mathrm{A} \rightarrow D_a D_a \mathrm{A}$ are already valid, the combination of $D_a \mathrm{A} \rightarrow I_a D_a \mathrm{A}$ with the former or $D_a \mathrm{A} \rightarrow D_a I_a \mathrm{A}$ with the latter yields the validity of $D_a A \rightarrow D_a D_a A$, which is exactly what had to be avoided. It is therefore easy to conclude that both axioms are at odds with the idea that meta-data actually do add something not yet present in the primary data itself.

## 5. Concluding Remarks

To conclude this article, two specific virtues of its general methodology deserve to be highlighted. Both concern the meaning and function of iterated *D*-modalities. The first benefit derives from the specific argument that was presented for the invalidity of $D_a A \rightarrow D_a D_a A$. When compared to standard rejections of the intuitively related KK-thesis, it should be noted that no appeal was made to either computational concerns or the higher standards required for *reflective* states. Instead, only the assumption of the added value of meta-data had to be introduced. In short: if meta-data were free in the sense of being obtainable by logical means only, it would be of no value at all. This

relies on a typical informational-theoretical concern which connects cost or value to informational content. If content, value, and logical consequence are so related, free-meta data are in fact informationally empty meta-data, and therefore only meta-data by name.

The second benefit derives from the interaction of the rejection of free meta-data and the failure of introspection for being informed. For starters, the rejection of $D_a A \rightarrow D_a D_a A$ is motivated independently from the rejection of $I_a A \rightarrow I_a I_a A$ in a system where $I$ is a **KTB**-modality. This is important in view of the fact that the free meta-data principle need not to be rejected in order to invalidate introspection for $I$. Instead, a logic for data which does not validate $D_a A \rightarrow D_a D_a A$ combines more elegantly with a logic for being informed which does not validate $I_a A \rightarrow I_a I_a A$. To wit, combined with the unobjectionable $I_a A \rightarrow D_a A$, $I_a A \rightarrow I_a I_a A$ enforces the clearly unacceptable $I_a A \rightarrow D_a A \,\&\, D_a D_a A \,\&\, D_a D_a D_a A \,\&\, \ldots$). This suggests that the rejection of free meta-data is in a sense more basic than the rejection of introspection for being informed: introspection trivializes the value and hierarchy of meta-data. Free meta-data, to the contrary, does not yield introspection for being informed.

## References


[1] Floridi, L., 2005, Semantic Conceptions of Information, *Stanford Encyclopedia of Information*, Zalta, E. N., ed., Stanford.
[2] Ackoff, R. L., 1989, From data to wisdom, *Journal of Applied Systems Analysis*, 16:3-9.
[3] Fetzer, J. H., 2004, Information: Does it Have To Be True?, *Minds & Machines*, 14(2): 223-229.
[4] Fetzer, J. H., 2004, Disinformation: The Use of False Information, *Minds & Machines*, 14(2): 231-240.
[5] Floridi, L., 2005, Is Information Meaningful Data?, *Philosophy and Phenomenological Research*, 70(2): 351-370.
[6] Sequoiah-Grayson, S., 2007, The Metaphilosophy of Information, *Minds & Machines*, 17(3): 331-344.
[7] Dretske, F., 1999, *Knowledge and The Flow of Information*. CSLI, Stanford.
[8] Williamson, T., 2000, *Knowledge and Its Limits*. Oxford University Press, Oxford.
[9] Landauer, R., 1996, The physical nature of information, *Physics Letters A*, 217(4-5): 188-193.
[10] Gochet, P. and P. Gribomont, 2006, Epistemic Logic, in: *Handbook of the History of Logic Vol. 6*, Gabbay, D. M. and J. Woods, ed., Elsevier.
[11] Van der Hoek, W. and J.-J. C. Meyer, 1995, *Epistemic Logic for AI and Computer Science*. Cambridge University Press, Cambridge.
[12] Floridi, L., 2006, The Logic of 'Being Informed', *Logique & Analyse*, 49(196): 433-460.
[13] Rowley, J., 2007, The wisdom hierarchy: representations of the DIKW hierarchy, *Journal of Information Science*, 33(2): 175.
[14] Fagin, R., J. Y. Halpern, Y. Moses and M. Y. Vardi, 1995, *Reasoning About Knowledge*. MIT Press, Cambridge / London.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# The Computer as Cognitive Artifact and Simulator of Worlds

Philip BREY
*University of Twente, The Netherlands*

**Abstract.** In this essay, the relation between computers and their human users will be analyzed from a philosophical point of view. I will argue that there are at least two philosophically interesting relationships between humans and computers: functional and phenomenal relationships. I will first analyze the functional relationship between computers and humans. In doing this, I will abstract from ordinary functions of computers, such as word processor, information provider, and gaming device, to arrive at a generalized account of the functional relationship between humans and computers. Next, I will explore the phenomenal relationship between humans and computers, which is the way in which computers transform our experience of and interaction with our environment or world. Both analyses, I will argue, point to a dual role of computers for humans: a cognitive role, in which the computer functions as a cognitive device that extends or supplements human cognition, and an ambient role, in which the computer functions as a simulation device that simulates objects and environments.

**Keywords.** Cognitive artifacts, extended mind, human–computer interaction, virtuality, simulation, functions, phenomenology, distributed cognition

## Introduction

How can the relation between computers and their human users be understood from a philosophical point of view? In this essay, I will answer this question from two angles. I will first explore the *functional relationship* between computers and humans. Computers have many functions for human beings, for example as word processor, calculator, information provider, and gaming device. What I will attempt in this essay is to abstract from these diverse functions and to arrive at an analysis of the functional relationship between humans and computers that is both more general and more profound. Next, I will explore the *phenomenal relationship* between humans and computers, which is the way in which computers transform our experience of and interaction with our environment or world. Both analyses, I will argue, similarly point to a dual role that computers play for human beings: a cognitive role, in which the computer functions as a cognitive device that extends or supplements human cognition,

and an ambient role, in which the computer functions as a simulation device that simulates objects and environments.

The distinction between functional and phenomenal relationships can be observed in previous work in the philosophy of technology. Thinkers like Ernst Kapp [1], Marshall McLuhan [2] and Arnold Gehlen [3], have tried to understand the *functional* relationship of technological artifacts to human beings. They did more than analyze the functions or functional capacities that various artifacts have. They also attempted to analyze how the functions of technological artifacts related to abilities already possessed by human beings, and the way in which artifacts build on, augment or extend these abilities. That is, they were specifically interested in functional *relationships* between artifacts and human beings rather than mere functions of artifacts. The *phenomenal* relationship between a human beings and a technological artifact concerns the way in which this artifact transforms his or her experience of and engagement with his or her world. Early phenomenological perspectives on human-technology relationships have been advanced by Merleau-Ponty [4] and Heidegger [5]. The most important contemporary phenomenological studies of human-technology relationships are those of Don Ihde [6,7].

In the next three sections, I intend to investigate how computer technologies maintain a functional relationship with their human users, and how this relationship is different from that of other technologies.[1] In the section thereafter, I intent to analyze phenomenal relationship between computer systems and human beings. The results of my study of functional relationships will be important building blocks for this phenomenal analysis. These two analyses are intended to clarify the fundamental roles that computers have in our personal lives, as functional items and as items that change the way in which the world is experienced and engaged.

## 1. Cognitive Artifacts

In an earlier paper, titled "Theories of Technology as Extension of Human Faculties," [9] I have argued that technological artifacts often serve to extend or augment existing human capacities or faculties. For example, instruments like microscopes and telescopes extend our vision, so that we can perceive objects or patterns that we could otherwise not perceive. Vehicles like bicycles and automobiles extend our locomotive abilities, so that we can move faster or with less effort. Tools like hammers and drills extend the ability of our hands to modify materials. Walls, heaters and air conditioners extend the thermoregulatory capacities of the human organism. Millions of other artifacts likewise extend perceptual, motor and regulatory functions of the human organism. Does computer technology likewise extend one or more of our faculties? According to Marshall McLuhan [2], it does. As I pointed out in my paper, McLuhan claimed in his *Understanding Media* that with the advent of electric media, it is no longer just perception and motor functions of humans that are extended by technology. He argued that electric media extend the information processing functions of the central nervous system, taking over functions of information management, storage and retrieval normally performed by the central nervous system. He specifically argued that digital computers extend creative cognition and higher thought. McLuhan hence saw

[1] These sections are based on Brey [8] in which the functional relationship between humans and computers is discussed more extensively.

the digital computer as extending cognition, as opposed to perception or motor functions.

I here intend to develop McLuhan's idea that the computer extends human cognition by building on human cognitive capacities. My focus will be on the question *how* computers extend human cognition, which I intend to answer by analyzing the functional relation between human cognition and computer activity. I will be arguing that the computer is a special kind of cognitive artifact that is capable of extending a broad range of cognitive abilities of human beings. The notion of a cognitive artifact has been introduced by psychologist Donald Norman [10]. According to Norman, there is a special class of artifacts that are distinguished by their ability to represent, store, retrieve and manipulate information. Norman calls such artifacts cognitive artifacts. He defines them as artificial devices designed to maintain, display, or operate upon information in order to serve a representational function. The keywords in this definition are "information" and "representation." They distinguish cognitive artifacts from other artifacts.

Norman's definition provides a clear criterion from distinguishing cognitive artifacts, such as thermometers, newspapers, clocks and Internet search engines, from noncognitive artifacts, such as hammers and automobiles. A thermometer has as its function is to inform us about temperatures. A newspaper has been made to store and display information on current events. A clock has been designed to accurately represent and display the time. An Internet search engine has been designed helps us to find information on the Internet. All these functions are representational functions. A hammer, in contrast, does not normally serve a representational function, as it does not normally maintain, display or operate upon information. There are perhaps some peripheral ways in which it may still serve representational functions. For example, it may contain a symbol or language that informs who the manufacturer is. And it may be put on a coffee table at home to remind oneself about a carpentry job that needs finishing. In that case it serves a representational function by making an indexical reference to the carpentry job. But it is not designed for such a purpose and therefore these cognitive functions are peripheral to its primary functions which are to hit nails and flatten or shape materials. Hence, it is not a cognitive artifact. Similarly, an architectural sketch that has been made to accurately represent a building is a cognitive artifact, whereas an artistic drawing of a nonexistent building is not a cognitive artifact, because it has not been designed to display information, but rather to please aesthetically.

Cognitive artifacts are properly called 'cognitive' because they, in quite straightforward ways, extend human cognition. They help us think, plan, solve, calculate, measure, know, categorize, identify, or remember. Various classes of cognitive artifacts may be distinguished, based on the primary cognitive capacity or capacities that they extend or aid. I will now list various basic cognitive abilities that have been recognized by cognitive psychologists, and illustrate how cognitive artifacts may extend or aid such abilities.

### *1.1. Memory*

Human memory is the psychological faculty by which we store information and retrieve it for later use. Cognitive artifacts that extend memory functions may be called *memory devices*. They are artifacts that help us encode, store and retrieve information. Sometimes, memory devices merely help us to locate information in our own memory.

For example, some banks issue cards that help you to reconstruct the PIN-code of your ATM card based on an easier to remember verbal code. More often, memory devices serve as memory systems themselves: they store information in organized ways. If memory is a means for encoding, storing and retrieving information, then any device which has this as one of its primary functions is a memory device. So a notepad is a memory device, as its function is to store notes for ourselves or others, and pens and pencils are memory devices used for inscribing data into external memory.

Psychologist Merlin Donald [11] has argued that one of the most important changes in the transition from Neolithic to modern culture is the emergence of a system of external memory storage, of which the storage of symbolic (linguistic) information is the most important. He claims that nowadays this external memory system contains more information than biological memories do, and that most human beings rely on it extensively. Media used for external memory storage include books, newspapers, microfilms, digital storage media, and others. For inscribing or reading them we have pens, pencils, microfiche readers, monitors and the like. Most important are paper and electronic (especially digital) storage devices. External memory devices serve in straightforward ways as extensions of human biological memory.

### *1.2. Interpretation*

Interpretation is also a fundamental human cognitive ability. Interpretation is the ability to assign meanings to input data, through the assignment of one or more concepts or categories. For example, when one tries to recognize objects in one's environment, one may perceive certain shapes and colors. To recognize what these shapes and colors stand for, one needs to apply concepts to them that make a 'fit'. For example, a curved yellow shape can only be recognized as a banana when the concept of a banana is applied to it. The interpretation of perceptual data is the way in which perceptual stimuli are made useful as objects of conceptual thought, which does not range over sensory images, but over concepts.

Interpretation can be qualitative or quantitative. *Quantitative* interpretation is the assignment of a numerical value to a perceived quality. Another word for this is *measurement*. Measurement is a cognitive activity that we typically, though not invariably, perform with the aid of artifacts, *measuring devices*, like thermometers, spectrometers, clocks, yardsticks, sextants, etc. The history of science and technology, if not economics, politics and management, is to a large extend a history of measurement, along with the measuring devices that have been developed for it. Measuring devices extend our abilities to estimate the size, number or intensity of phenomena in the world, and are hence extensions of our ability to interpret the world.

*Qualitative* interpretation is the assignment of a qualitative concept or category to data. There are many cognitive artifacts that aid in the qualitative interpretation by giving criteria, templates or examples for the application of a concept. For example, color charts aid in the correct identification of colors. A book on animal tracks, with drawings op typical animal tracks, helps one in the identification of tracks observed in the woods. Medical texts list criteria for the correct identification of diseases. Few artifacts exist, however, that do not just support qualitative interpretation but that do the interpretive work themselves. The digital computer is an artifact capable of autonomous interpretation. Most qualitative interpretation performed by computers takes symbolic inputs, such as sentences, numbers or names, and assigns categories to them. For example, a computer program may take names of animals and classify them

as "reptile," "mammal," "bird," "amphibian," etc. Or it may take a sentence, and parse it by assigning grammatical roles to words and phrases. Computers are also capable, when suitably programmed, to recognize objects and scenes in pictures, although their capabilities to do this are more limited.

### *1.3. Search*

When we interact with the world, we often actively look for things that we are trying to locate but have not observed yet. We constantly look around for people, pens, purses, stores, food, stamps, road signs, words, barcodes, and numerous other things that we need to see, locate or use. The ability to search and subsequently recognize things is one of our fundamental cognitive abilities. Searches sometimes take place with exact specifications of what you are looking for, but more often they are heuristic, and take place according to hypotheses: you assume that there is something in your vicinity that meets a set of loosely formulated criteria, and search for something that meets these criteria. Searches do not just take place in the external world; we also frequently search our own memories for information.

Search is a cognitive process, because it involves activities like mental scanning and pattern matching. It is another process that can be assisted by cognitive artifacts. Cognitive artifacts can aid search by structuring the search space in such a way that it can be more easily scanned, and by 'flagging' types of items that one may scan for (e.g., by marking them with colors or symbols). Examples of cognitive artifacts that aid search are labels and filing systems. A special ability of computer systems is that they can perform searches themselves. They can do so because of their ability to do pattern matching and their ability to systematically scan through a search space.

### *1.4. Conceptual thought*

The most important cognitive ability that distinguishes human cognition from animal cognition is the ability to engage in conceptual thought, and particularly the ability to engage in abstract thought, using abstract concepts. Conceptual thought is the ability to arrive at new conceptual structures (ideas or beliefs) through the modification (analysis or synthesis) of existing ones. Conceptual thought often involves *problem solving*: it often involves cognitive goals like finding the solution to a mathematical equation, determining the best way to furnish a room, finding an adequate translation into English for a sentence in Spanish, or thinking up the most diplomatic answer to a potentially embarrassing question. Problem solving can be aided by cognitive artifacts that help to arrive at an accurate representation of the problem space or of the kinds of steps to take to find a solution, such as models and diagrams, and procedural manuals. Computer systems are, again, special in that they are capable of autonomous problem solving. When suitably programmed, computers are capable of solving equations, thinking up room designs, translating sentences from Spanish to English, or answering questions. Computer intelligence of course still has its limitations. Results are not impressive, for example, in the areas of language use and reasoning in informal domains. Nevertheless, computers are nowadays frequently used for all kinds of tasks that ordinarily require conceptual thought, whether they are performing calculations, correcting grammar, engaging in dialogue, planning distribution routes, or designing copying machines.

## 2. Computer Systems as Cognitive Artifacts

Among the many cognitive artifacts that exist, computer systems are certainly unique. As has been observed in the previous section, computers are special in that they often go beyond the role of facilitating or aiding human cognition: computers are capable of performing cognitive tasks *autonomously*. Computers are special because they are capable of *actively manipulating representations*. Most other cognitive artifacts cannot manipulate representations, because they are not capable of systematically discriminating different kinds or representations and responding to them in meaningful ways. More specifically, still, computers are *physical symbol systems* [12], systems that manipulate (meaningful) physical symbols in virtue of their formal (syntactical) properties according to a finite set of operations, thus producing as output (meaningful) symbolical structures. This capability is the reason that computer systems are the most versatile and powerful cognitive artifact that can either support or perform almost any cognitive task.

The functional relation that computers, as cognitive artifacts, have to their human users is hence that they extend cognition. Specifically, they extend the memory, interpretation, search, pattern matching and higher-order cognitive abilities of human beings. There is not, however, a single way in which computer systems functionally extend human cognition. I observed that computer are capable of autonomous cognitive processes. But they may also serve as a mere facilitator of human cognitive processes, as happens for example in word processing. I will now go on to further analyze how exactly computer systems add to, augment or replace the cognitive capacities of human beings.

As my point of departure, I will take a set of distinctions made in the formerly mentioned essay "Theories of Technology as Extension of Human Faculties."[9] In this essay, I argued that artifacts that amplify the functioning of human organs may maintain three different types of relations with these organs. An artifact may *replace* the functioning of an organ by performing the function of that organ in a way that makes the organ redundant. For example, when driving a car, one's legs are not used as a means for transportation. An artifact may also *supplement* an organ that it extends, by performing a function that the organ in question is also performing. For example, clothing adds to the protective and temperature control functions already performed by the skin. Third, an artifact may *enhance* the functional powers of the organ that it extends, not by independently performing a function that resembles the organ's function, but by cooperating with the organ in a way that enhances its activities, in this way engaging in a *symbiotic relationship* with the organ. For example, a telescope extends visual perception by teaming up with the eye to form a new functional unit consisting of telescope-plus-eye that is capable of doing things that neither the telescope nor the eye is capable of doing by itself.

The relevant faculty or 'organ' that is extended by computer systems is our faculty of cognition, located, according to neuroscience, in our brain, specifically in the neocortex. Is a computer system an artifact that mostly replaces, supplements or enhances human cognition? All three roles are visible in computer systems. In its early days, the computer was often called the 'electronic brain,' and a common fear was that computers would replace human brains as the primary locus of cognitive activity. The computer as a replacement of human cognition is an autonomous information processing system that operates like a human cognitive agent, producing its own plans, solutions, and other knowledge structures without human intervention. In this role, the

computer fits the early ideals of artificial intelligence research to 'build a person,' and the ideal of expert systems research to replace human experts.

The idea of the computer as a replacement of the human cognitive system has never been fully realized, and it is nowadays recognized that AI's dream to 'build a person' still depends on significant breakthroughs in AI research that have not been realized in past decades. The idea of the computer as a supplement to human cognition, in contrast, was already powerful in the early days of the computer and this idea still holds currency today. The computer in its supplementary role does autonomous information processing, but remains limited to those tasks that are tedious, time-consuming, or error-prone when performed by humans. These are tasks like doing large calculations ('number crunching'), database searches, and organizing and reformatting data. The implicit distribution of labor between humans and computers is then that humans perform the more intuitive and creative cognitive tasks and are responsible for the overall structure and goals of large cognitive tasks, whereas computer systems autonomously perform more tedious or time-consuming cognitive tasks that are defined as 'subroutines' within such larger cognitive tasks.

Since the rise of the personal computer, however, a third powerful interpretation of the role of the computer has emerged: that of a versatile tool that we handle directly and that enhances our own power to get work done. In this role, the computer is not an autonomous cognitive unit, but a cognitive aide, that enhances our own cognitive powers. It does not perform cognitive tasks by itself, but helps us to perform them. Our relation with the computer in this role is more *symbiotic*: the performance of a cognitive task depends on the information-processing abilities of both human and computer, and the exchange of information between them. When we use a word processor, spreadsheets, web browsers, and other software tools, the cognitive tasks we perform, such as producing well-formatted documents, performing calculations, or navigating the Web, are performed in cooperation with the computer. When we check the spelling of a document with the aid of a spelling checker, for example, this cognitive task depends on both the ability of the spelling checker to identify possible misspellings, and our own ability to operate the spelling checker and to decide whether its proposed corrections are valid.

Even in their role as tool, however, computers still engage in autonomous information processing. The previously mentioned spelling checker may not autonomously correct the spelling of a document, but it does make autonomous proposals. On the other hand, the computer in its role as a supplement to human cognition still requires a knowledgeable human operator, so its operations are not entirely autonomous. So the distinction between supplementary and enhancement roles of the computer is by no means absolute. In both cases, cognition is made into a distributed process that depends on the information-processing abilities of both humans and computers. The mutual dependency is greatest, however, when the computer functions as an enhancement of human cognition. In these cases, the computer operates *in tandem* with the human mind, and the integration of cognitive functions becomes so great that human and computer are best regarded as a single cognitive unit, a *hybrid cognitive system* that is part human, part artificial, in which two semi-autonomous information-processing systems cooperate in performing cognitive tasks.

## 3. Computing and World-Simulation

In its early days, roughly from the late forties to the late seventies, the computer was exclusively a cognitive tool, since the only tasks that it was designed to do were cognitive tasks, like calculation and information management. This has changed with the advent of computers with good graphical and multimedia abilities in the late seventies, eighties and nineties. These computers, most of them personal computers, acquired new functions that were not primarily cognitive. When a computer system is used to make a creative drawing, to play an adventure game, or to listen to music, it is not used as a cognitive artifact, because the performed functions are not information functions: artistic drawings, adventure games and music are not meant to inform, but rather to please or entertain. These activities may involve cognitive activity (almost any activity does), but their principal goals are not cognitive. The computer systems and software that supports such activities therefore do not qualify as cognitive artifacts.

Most of these new noncognitive functions of computer systems critically depend on newly acquired abilities of such systems to graphically represent, simulate or model interactive objects, structures and environments. I will term such abilities *simulation abilities*. With the rise of high-quality graphical capabilities in computers, the computer is no longer just a *cognitive device*, it is now also a *simulation device* (cf. Turkle[13]). To wit, the two functions, cognition and simulation, are not mutually exclusive. In fact, many of the early efforts at graphical simulation were aimed at making the computer a better cognitive artifact. The Xerox Star was, in the late seventies, the first computer to make use of a graphical user interface, using a desktop metaphor that has since been copied by Apple (Macintosh) and by Microsoft (Windows). Desktop interfaces offer graphical user environment with documents, folders, trash cans, rulers, pencils and in and out boxes, that can be operated and manipulated in ways not unlike their physical counterparts. Their primary function, however, is to better support information processing activities, particularly those performed in offices.

The advantage of graphical user interfaces over the older symbol-based interfaces (such as DOS and UNIX) is that they rely on our sensorimotor abilities to orient ourselves in space and to recognize and manipulate objects. Symbolical user interfaces make no good use of our sensorimotor abilities, and instead rely on our capacities for abstract thought. However, because people's sensorimotor abilities are usually better developed than their capacity for abstract thought, it pays to treat data and programs as manipulable, visible objects, when possible. As a result, the tendency in software development has been to devise programs in which data strings, (sub)programs and procedures are translated into visual icons and actions like clicking, 'dragging,' and scrolling.

Around the same time that graphical user interfaces came into vogue, the first noncognitive graphical computer applications started to become popular: graphical computer games and creative software like paint and music programs. These applications are noncognitive because they do not have as their primary function to assist in the performance of information-processing tasks. Instead, they are intended to extend our means for entertainment and creative expression. They do so by simulating physical environments, objects and events. Tools are simulated with which we can interact with the world, like paint brushes, golf clubs, wrenches and guns, and the objects encountered in the graphical environment can be programmed to respond visually and aurally like their physical equivalents. Many environments can even be navigated, representing a position for us, and giving us the option to move to a different

position. And in some environments, we can even interact verbally or nonverbally with computer-generated characters.

The computer in its role as (graphical) simulation device functions perhaps less as an extension of ourselves than as an extension of our world. The virtual interactive environments generated by computers offer us new structures to experience, navigate and interact with. They are hence an augmentation of the world as it existed before. Although these structures are not physically real, they are nevertheless meaningful or useful to us, sometimes as much as their physical equivalents. They can clearly be useful for performing cognitive tasks, as I argued in my discussion of graphical user interfaces. They are also useful for learning, particularly for learning sensorimotor skills, through their ability to faithfully simulate physical structures that we interact with. And they are useful for entertainment and creative activity. They hence serve a functional role as broad and diverse as the functional roles of many of the structures encountered in the physical world.

## 4. A phenomenology of human-computer relationships

In the previous three sections, I analyzed the functional relationships that exist between computer systems and their human users. An analysis of functional relationships reveals fundamental functional roles of computers and their functional relation to the abilities of their users. In this section, I will analyze phenomenal relationships. The phenomenal relationship between a human being and a technological artifact is given by the way in which this artifact transforms his or her experience of and engagement with his or her world. Such an analysis can reveal significant shifts in the way people experience and engage with their world. Understanding such shifts is necessary for an understanding of the significant social, cultural and psychological changes that accompany the digital revolution.

Philosopher Don Ihde [6,7] has argued that, from a phenomenal point of view, there are three basic types of relations between technological artifacts and human beings. First, they may mediate between humans and the world. This happens in *mediation relations*. In mediation relations, an artifact is a means by or through which the world is experienced, engaged, or studied. In this way, the artifact becomes part of one's intentional stance towards the world, and co-determines the way in which the world is experienced or engaged. Ihde recognizes two basic mediation relations. In *embodiment relations*, an artifact mediates between self and world through a direct mediation of perception of or behavior towards our world (cf. Brey [14]). Embodiment relations occur in the use of artifacts like glasses, telescopes, bicycles and hammers. When used, such instruments become transparent means through which the world is perceived or acted on. They are 'incorporated' into our perceptual apparatus and motor programs, and may even come to feel like they are part of us. In hermeneutic relations, an artifact mediates between self and world through one or more representations of the world. Hermeneutic relations occur in the use of artifacts like maps, control panels and thermometers. These are artifacts that mediate our experience of or engagement with the world through symbolic or pictorial representations of the world.

The second type of relation identified by Ihde, next to the mediation relation, is the *alterity relation*. In alterity relations, artifacts are experienced and engaged as objects encountered in the world. In alterity relations, we direct our attention to an artifact, and observe it or interact with it. Alterity relations may involve attitudes and feelings

towards the artifact, for example of admiration, resentment or love. Any artifact may take part in alterity relations, although some, like automobiles, sculptures and jewelry, receive more attention than the average artifact. The third type of relation is the *background relation*, in which artifacts function as background objects in the world that people do not directly engage with, but that constitute part of the context in which people operate. Examples are central heating systems and electric lights that function as background objects that affect the way we experience and interact with the world without functioning either as a medium or as an object of perception or engagement.

All four of these relations, embodiment, hermeneutical, alterity and background, can be established between human beings and computer systems. Most salient, perhaps, are hermeneutical relations. When computers are used as cognitive artifacts, a hermeneutical relation is established with them. In such a relation, the computer represents information about the world, and the user reads, observes, navigates, or reformats this information, and may also add information about the world herself, or instruct the computer to produce or transform information. A consequence of the frequent use that is made of computers as a hermeneutic device is that increasingly, our knowledge of the world is mediated by computers. Importantly, this mediation is not a passive process of rendering pictures or texts. Computer systems also engage in interpretation, calculation, reasoning, planning, and decision making, and thus play a very active role in the formation of our knowledge of the world and our plans for acting on it.

Also salient in our use of computers are alterity relations. The computer is frequently experienced and engaged with as an 'other,' an artifact that is interesting both to observe and to interact with. The alterity relations that people establish with computers are more complex and emotional than those established with most other artifacts. As Sherry Turkle [13,15] and Reeves & Nass [16] have observed, computers are often anthropomorphized: they are considered and treated as persons. Anthropomorphization also occurs with other artifacts, for example with automobiles. But computers have much greater similarities with human agents: they autonomously perform actions, they like other human beings respond to symbols, to human language even, and they appear capable of intelligent behavior. Whereas much of this behavior is attributed to 'the computer,' there is an increasing tendency to model agents *on* the computer. That is, agents capable of performing specific tasks are identified on the computer, like office assistants, search agents, Tamagotchis and chess opponents, and may even be equipped with expressive faces and speaking voices. In this way, our world is increasingly populated with artificial agents to which we establish alterity relations, whether cooperative, competitive or neutral. As Turkle [15] has argued, a consequence of our experiences with the computer as an 'other' that is located somewhere in between a dumb machine and an intelligent organism is that people are renegotiating their conceptions of intelligence, mind, self, and life, along with their attitudes. Is a Tamagotchi (a simulated pet) really alive, and if not, why can its death feel so real? Is the display of intelligent behavior a sufficient condition for having a 'mind'? These are the kinds of issues that have to be (re)negotiated in a world filled with artificial agents.

Whenever we use a computer, we also establish embodiment relations with it. To be precise, we establish embodiment relations with the input and output devices, like keyboard, mouse, joystick and monitor. But through these physical devices, we are also capable of establishing embodiment relations with virtual objects on the screen, like pointers and paint brushes (e.g., in a paint program) and guns, telescopes and

automobiles (e.g., in three-dimensional simulations and games). Such embodiment relations often exist simultaneously with other relations, like hermeneutical and alterity relations. For example, when someone plays chess against a computer opponent, both embodiment, hermeneutical and alterity relations are established. The special character of embodiment relations with computers is that the 'world' that is experienced and engaged through their input and output devices and their virtual correlates is not a physical but a virtual world. This implies that sensorimotor skills are learned that relate not to physical objects and environments, but to virtual ones, with all their peculiarities.

Let us finally turn to as phenomenon that is difficult to reconcile with Ihde's typology of mediation, alterity and background relations. I argued in the previous section that virtual environments generated by computers may be regarded as extensions of our world that offer us new structures and environments to engage. Such simulations are like physical worlds, in that they can be interacted with and navigated, and the objects they contain can relate to us in all the ways indentified by Ihde: embodiment (e.g., a virtual wrench), hermeneutical (e.g., a virtual map), alterity (e.g., a computer-generated dog) and background (e.g. a virtual light source). One way in which virtual environments may be understood in the context of Ihde's scheme is as aspects of computer systems to which alterity relations are established. In this view, a virtual environment is a computer-generated artifact that we experience and interact with, as in an alterity relation. Only, the structure is so rich that within the context of this alterity relation, we can establish more specific alterity relations with substructures of the environment, as well as embodiment, hermeneutical and background relations with yet other substructures.

Alternatively, virtual environments may be analyzed as straightforward extensions of the physical world that should not be understood as complex artifacts but as *worlds*. Worlds are not artifacts to which we have relations, but contexts within which such relations are established with specific objects. Both interpretations of virtual environments, I wish to suggest, have their own worth. The first interpretation downplays the idea of virtual environments as genuine worlds and tells us that the long hours we spend designing virtual landscapes, playing Tomb Raider or surfing the Internet are not hours spent in the real world but with a machine. The second interpretation accepts virtual environments as genuine worlds, and hence the idea that our experience and interactions in virtual environments can be as meaningful and 'real' as those in the physical world. Both interpretations are valid because virtual environment are ambiguous in precisely this way: they are both mere artifacts and genuine worlds, depending on how much one invests and 'believes' in them.

## 5. Conclusion

The functional analysis of computer systems presented here has identified computer systems as both cognitive devices and simulation devices. In its role of a cognitive device, the computer extends human cognitive faculties by both supplementing and enhancing them. It is particularly in this latter role that collaboration between human mind and computer system becomes so close that one may speak of a hybrid cognitive system that is part human, part artificial. In its role of a simulation device, the computer does not so much extend human faculties as extend the world. Computer-

generated, virtual environments offer extensions of the physical world that are useful for entertainment, creative activity, learning and social interaction.

The phenomenal analysis of computer systems has revealed that computer systems, in their role of cognitive device, engage with their users in hermeneutical and alterity relations. It was claimed that a consequence of the many hermeneutical roles played by computer systems is that our knowledge of the world is increasingly mediated by computers, not just passively, but actively, in that computers are actively engaged in the production of knowledge structures. A consequence of the alterity roles played by computers is a blurring of the traditional distinction between (intelligent) human beings and (dumb) machines, so that many of our concepts and attitudes regarding this distinction need to be reevaluated. Finally, it was claimed that virtual environments can be interpreted either as computer-generated artifacts that are a mere object of alterity relations, or as genuine worlds, that can be navigated and interacted with in myriad ways. It was argued that both interpretations have their validity, as their ambiguity between artifact and world is an essential property of virtual environments.[2]

## References

[1] Kapp, E. (1877). *Grundlinien einer Philosophie der Technik: Zur Entstehungsgeschichte der Cultur aus Neuen Gesichtpunkten.* Braunschweig: Westermann.
[2] McLuhan, M. (1966/1964). *Understanding Media: The Extensions of Man*. New York: McGraw-Hill. Paperback edition, 1966.
[3] Gehlen, A. (1980/1957). *Man in the Age of Technology*. P. Lipscomb (trans.). New York: Columbia University Press. Originally published in German as *Die Seele im Technischen Zeitalter: Sozialpsychologische Probleme in der Industriellen Gesellschaft.*
[4] Merleau-Ponty, M. (1962/1945). *Phenomenology of Perception*. C. Smith (trans.). New York and London: Routledge. Originally published in French as *Phénoménologie de la Perception*.
[5] Heidegger, M. (1962/1927). *Being and Time*. J. Macquarrie and E. Robinson (trans.). New York: Harper and Row. Originally published in German as *Sein und Zeit*.
[6] Ihde, D. (1979). *Technics and Praxis: A Philosophy of Technology*. Boston: D. Reidel.
[7] Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Bloomington: Indiana University Press.
[8] Brey, P. (2005). 'The Epistemology and Ontology of Human-Computer Interaction,' *Minds and Machines 15*(3-4), 383-398.
[9] Brey, P. (2000a). 'Technology as Extension of Human Faculties.' *Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology, vol 19*. Ed. C. Mitcham. London: Elsevier/JAI Press.
[10] Norman, D. (1993). *Things that Make us Smart: Defending Human Attributes in the Age of the Machine*. Reading, MA: Addison-Wesley.
[11] Donald, M. (1991). *Origins of the Modern Mind.* Cambridge and London: Harvard University Press.
[12] Newell, A. and Simon, H. (1976). 'Computer Science as Empirical Inquiry: Symbols and Search,' *Communications of the Association of Computing Machinery*, 19: 113-126.
[13] Turkle, S. (1995). *Life on the Screen; Identity in the Age of the Internet.* New York: Simon & Schuster.
[14] Brey, P. (2000b). 'Technology and Embodiment in Ihde and Merleau-Ponty.' *Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology, vol 19*. ed. C. Mitcham. London: Elsevier/JAI Press.
[15] Turkle, S. (1984). *The Second Self. Computers and the Human Spirit*. London: Granada.
[16] Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and new Media Like Real People and Places*. CSLI Publications.

[2] Previous versions of this paper were presented at E-CAP 2005, Sweden, Delft University of Technology, The Netherlands, and Virginia Institute of Technology, USA. I wish to thank the commentators at these events as well as two anonymous referees for their comments. I also wish to thank Maurice Liebregt for his great help in editing this essay.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# The Panic Room: On Synthetic Emotions

Jordi VALLVERDÚ [a, 1] and David CASACUBERTA[b]
[a] *Philosophy Department, Universitat Autònoma de Barcelona*
[b] *Philosophy Department, Universitat Autònoma de Barcelona*

**Abstract.** Emotions and feelings are basic regulators of human activity. We consider that intelligence is an emergent property of systems and that emotions play a basic role within those systems. Our main aim is to create a system (called The Panic Room) based on a bottom-up "Ambient Intelligence" context which develops a proto-emotion of fear and pleasure in order to detect a dangerous event and react to it. The system labels the signals from the sensors which describe the surroundings as either negative or positive. Either option has a specific signal that is used to change the way further perceptual signals will be processed as well as generate possible behavioural responses to a possible danger. Responses are automatic and embedded (or *hardwired*) in the system.

**Keywords.** Synthetic emotions, ambient intelligence, affective, computing, bottom-up, simulation

## Introduction

Emotions and feelings are basic regulators of human activity. In fact, they are the basis of our interaction with the world: through pleasure, pain, hunger or fear, we create intentional dispositions, acting like homeostatic controls over our actions.

According to the traditional old-fashioned view of emotions, rationality and 'knowledge' should be located beyond emotions. Plato, Aristotle, Kant and Descartes have defended such an argument. According to the classical Western as well as Eastern philosophical traditions, the sage is the rational (and emotionless) man or woman who achieves this lack of emotion through the *ataraxia* or *boddhi*. In this sense, emotions are second level action inducers or, at least, a different and inferior human drive. As Blaise Pascal perfectly synthesised in *Pensées* (1660): The heart has its reasons, of which reason knows nothing. For Pascal, abstract reason is superior to the emotions although emotions have a special role in human existence (more closely related to the question of faith).

[1] Corresponding Author: Prof. Dr. Jordi Vallverdú, Philosophy Department, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Barcelona), Catalonia, Spain; E-mail: jordi.vallverdu@uab.cat.

Nevertheless, recent decades of scientific research in neurophysiology have shown how emotions are not just any other part of human activity, but a fundamental one [1]. Emotional states had been historically banned in the territory of human rationality (Descartes and his "res cogitans" was one of the more convincing proponents of this approach). Meanwhile evolutionary approaches to consciousness [2] or studies of the emotions ([3] affirms that emotions gave rise to consciousness) have shown that the origin of consciousness, lying in the structure of the nervous system (which enables data feedback loops, the cause of the emergence of consciousness), might be emotion (rather than perception) and that experienced sensations (i.e. *qualia*) inherently require someone to experience them [4].

Although this is not the space to talk of consciousness, but rather of emotions, we need to recognise that this is a complex and poorly understood scientific question. To mention just one example, *Scientific American* included in its October 2007 edition the open debate between two experts on consciousness, the neuroscientists Christof Knoch and Susan Greenfield. If consciousness is still a 'black box' for neuroscientists, it is absurd to spend our time trying to define or 'imitate' it within AI environments. We could argue similarly about various *qualia.*

Our point is that, although there is a good deal of evidence pointing to a connection between consciousness and emotions (the latter being necessary for self-perception and, therefore, consciousness), we don't need to be concerned with this fact. We are merely trying to design an artificial device able to learn from, and interact with, the world by using two basic information types: positive and negative. These can be considered as proto-emotions and, assuming we can establish this analogy with human emotions, we can emulate their usefulness in the fight for survival by creating helpful behavioural rules such as "this is harmful, don't touch it" or "this produces pleasure, eat it". Ours is not a cognitivistic simulated model of human emotions, but an experiment dealing with how to simplify the interactive possibilities between a single entity (artificially constructed, either physically or by means of computational simulation) and its environment, based on a system that we call 'the emotions', when we refer them to human beings. We are developing the first steps towards an evolutionary machine, defining the key elements involved in the development of complex actions (that is, creating a *physical intuitive ontology*, from a bottom-up approach).

At the same time, there are several kinds of studies of emotions in synthetic environments, such as affective computing [5] or sociable robots [6]. Some authors have also tried to develop computational models of artificial emotions [7-11] or have drawn attention to the interesting phenomenon of emotions within artificial environments [12-14].

Our intention is not to reproduce human emotions in a computational model, but to develop an approach to synthetic emotions from the ground up. Like [15], we consider that intelligence is an emergent property of systems and that emotions play a basic role within those systems [16,17]. In order to achieve an 'artificial self' we must not only develop the intelligent characteristics of human beings but also their emotional disposition towards the world. We are putting the artificial mind back into its (evolutionary) artificial nature.

## 1. What Are Artificial Emotions and Why Do We Need Them?

### *1.1. Defining 'emotion'*

From a basic perspective we use the ideas of [18], who define emotions, or *proto-emotions* (as we prefer to call them), thus:
We can use the idea of an alarm system to attempt a very general definition of "emotion": An organism is in an *emotional state* if it is in an episodic or dispositional state in which some part of it, whose biological function is to detect and respond to 'abnormal' states, has detected something and is either (1) *actually* (episodic) interrupting, preventing, disturbing or modulating one or more processes which were initiated or would have been initiated independently of this detection, or (2) *disposed* (under certain conditions) to interrupt, prevent, disturb, etc. such processes, but where such action is currently suppressed by a filter or priority mechanism.

The use of the term 'proto-emotions' is a straight choice that we make among the exact number of basic emotions, a controversial question among experts on the topic, a debate that we cannot develop here due to its extensive nature. We consider the existence of two activated signals: pain and pleasure (the so-called 'proto-emotions'), and one inactivation state (neutral). Both activation signals can escalate in intensity (*pain* leads to *pain+* and this leads to *panic*; *pleasure* leads to *happiness*).

### *1.2. Defining 'synthetic emotion'*

If we have stated that an emotion is like an alarm system, then we can acknowledge a concept from cybernetics: *feedback*. The interesting point is to see how an emotion is a homeostatic device that reacts in the particular way that is conditioned by its physical structure. Thus, an emotion is an integrated or embedded reaction system.

For our purposes, a 'synthetic emotion' is an independently embedded (or *hardwired*) self-regulating system that reacts to the diverse inputs that the system can collect (internal or external). This is the first layer of an artificial device's interaction with the world. Therefore, we consider an emotion as:

a) a signal, which is generated after evaluation of surroundings and which produces automatic responses;
b) a regulation system with valence (+ or -).

We have 2 basic emotions + and -, that we can term 'pleasure' and 'pain', respectively. We know that here there is a strong anthropomorphist analogy, however we wish to avoid falling into common misconceptions and false analogies about human-machines, and develop a degenerating research program [19]. In this regard we quote McCarthy [20]:

> To ascribe certain *beliefs*, *knowledge*, *free will*, *intentions*, *consciousness*, *abilities* or *wants* to a machine or computer program is *legitimate* when such an ascription expresses the same information about the machine that it expresses about a person. It is *useful* when the ascription helps us understand the structure of the machine, its past or future behaviour, or how to repair or improve it. It is perhaps never *logically required* even for humans, but expressing reasonably briefly what is actually known about the state of a machine in a particular situation may require ascribing mental qualities or qualities isomorphic to them.

**Table 1.**

| Proto-emotions | Related emotions | States |
|---|---|---|
| Pain | Pain + | Activated |
| | Panic | |
| Pleasure | Happiness | |
| - | | Neutral |

Emotional intentionality should be included under the categories explained by McCarthy, and this is what we have done with our research. Our initial purposes are to create a basic device, isomorphic to nervous systems that make possible a special kind of interaction with the world, according to their physical structure and survival strategies (social, mental, skills, specialisation...).

So, we are working with two emotions that create dynamical changes between two attitudinal states: calm (neutral) and activation (pain or pleasure). The first one is related to the homeostatic equilibrium of the whole system, factors such as ambient temperature and the necessary food supply. All systems tend to satisfy their necessities when they receive signals that ask for different solutions. The second one, pain or pleasure, concerns the basic alarm system of nervous systems to achieve a bodily answer. Of these two signal activators, we consider pain a more basic state and emotion: pain and the fear of suffering from it. Therefore, our system might be fearing something, or could be completely calm. The process to reach a calm state needs positive inputs, which could be defined as pleasure but they are mere stabiliser inputs towards calm states. In table 1. you can see these combinations:

And, therefore, the system will change states according to positive or negative signals (See figure 1).

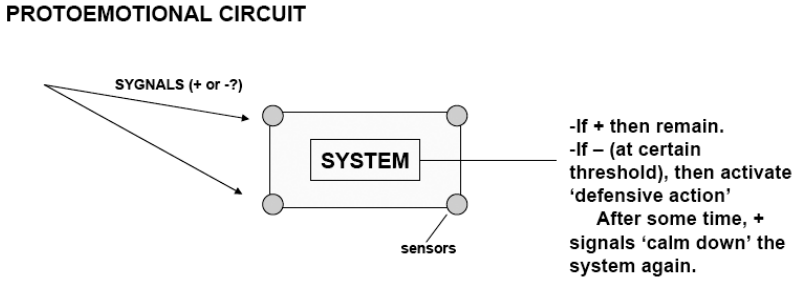



**Figure 1.**

*Defining our vocabulary, which represents the internal states of TPR:*

- *Pain*: TPR activates the pain signal when the visitor is close to the switches.
- *Pain+*: The visitor has activated the first switch and is close to the second one. This is a more intense pain.
- *Panic*: After activating the first two switches, the visitor is now close to the third. At this moment TPR closes the system and changes to automatic control.
- *Pleasure*: The visitor has activated one of the switches but has spent too much time in arriving at another one, so TPR is momentarily experiencing pleasure, because the visitor might try again to activate one switch. In our simulation the visitor has four seconds to move from one switch to the next. If he or she takes more time, the system is reset and he has to start all over again.
- *Happiness*: The visitor is absent and has not activated a panic situation. Therefore, there is no danger.

### *1.3. Why do we need them?*

If we want to create better rational artificial devices, we must create artificial emotions for several reasons. First of all, there is strong evidence from neurophysiology that shows us the deep relationships between emotions and rational processes. The fact is that without a good emotional structure, rationality is not possible. So there are fundamental emotional aspects of intelligence.

Secondly, there is the question of intentionality, completely necessary to avoid problems like the classic *frame problem* [21]. The best strategies to achieve a goal can be encoded previously into a machine through a program or they can be discovered by learning. Our idea is that the possibility of self-instruction is completely determined by the physical structure of the artificial device, which requires artificial emotions in order to have a teleological orientation towards the world (a deep and structural orientation). If we use humans as an example, we know that humans have not only their genetic and cultural programming (different but complementary code layers for action) but also a biological structure, all of which act as a filter and as an orientation towards the feedback between themselves and the world. Emotions have the role of controlling the interactions with the other objects and beings. Machines without (physically mediated) intentionality have no sense of complex thinking and being.

Thirdly, there is the empathy question. Recent studies have discovered the neural basis of learning by mirror neurons [22,4]. The emotional basis of empathy is something obvious and it makes possible self-instruction by imitation.

Thus rationality, intentionality and learning are basic characteristics of complex systems, such as humans, that are emotionally embedded [23,24]. Therefore, if we want to create better and smarter artificial devices, we must put emotions into them.

Some space should be devoted also to the philosophical relevance of this study. Our reasons for designing such an experimental device are threefold:

1. By developing a bottom-up approach to emotions we try to emphasise the need to change our philosophical models for understanding the role of emotions and their significance in contexts of decision making and rationality. So far we have indirect empirical evidence that emotions, in fact, are present in most cognitive processes which are considered “rational” in humans, as shown in [1], for example. But this doesn’t show, *per se*, that emotions are needed for rational action. After all, they

could be some sort of epiphenomenon with physiological causes. However, if we can show that 'emotions' radically improve the way a robot negotiates its movements within a certain space, as modelled in the Panic Room, we can postulate their role in the decision making processes that generate human actions within a space, thus we are building specific philosophical evidence for the importance of emotions.

2. Clearly, a large element of the debate about *qualia* and "the hard problem" relates in some way to emotions. Emotions are one of the basic mechanisms that allow us to generate some sort of *qualia* perception and traces of consciousness. It is the fact that we associate some emotions and pleasure and pain signals with certain perceptions that makes claims about the irreducibility of *qualia* somewhat plausible. By trying to define a system of synthetic emotions we are therefore opening up a novel and somewhat experimental way of understanding the relationships between emotions themselves, and between emotions and the actions they provoke - we are therefore tackling some of the "hard problem" related issues.

   Put another way, the reason why people find the idea of a computer having emotions problematic is akin to the difficulty of conceiving of a robot enjoying the sunset or a glass of wine. Affective computing (see [5]) does not tackle the issue, as everybody would admit that computers can easily have the cognitive part we associate with emotions. Like zombies in the famous paradox created by David Chalmers, we can have automata which are able to behave like human beings in every aspect, but do not have any of the feelings which we have. On the other hand, a bottom-up approach to synthetic emotions tries to imagine whether very simple feelings – such as pleasure and pain in the case of our system - could one day be incorporated into a computer.

   This also connects with the idea presented in former studies about the relationship between emotions and consciousness, as argued in [3] and [27]. Having consciousness of something (hearing, seeing, touching) clearly involves qualia. On the one hand, as argued in most literature about emotions, these mental states do not make sense without some qualia attached to them ([1], [28], [29]), therefore they are the perfect candidates to deal with *qualia.* On the other hand, as argued in [27], emotions are also very useful to connect basic physical and physiological states with higher cognitive states. Therefore, although details are still very sketchy and there is still a long way to go, some phenomenal states associated with certain types of consciousness, especially those related to perceptual states have clear connections with emotional states, and therefore it would be very useful to understand them better in order to have an explanation of how consciousness works. A bottom-up approach to synthetic emotions, then, may also turn out to be useful when looking for new approaches and tools to analyse what consciousness is and how it relates to perception and *qualia.*

3. As shown in [5] and [24] there is also a practical application of such studies. If – as the affective computing studies seem to show - computer algorithms can be improved by including some "emotional knowledge" within them, then it is also clear that, while a top-down approach is useful, a bottom-up one can also be of help. The application we present here could also have been solved without any recourse to emotions, since it is a very simple problem from a computational point

of view, but it nevertheless shows a different way to design and implement affective computing solutions.

## 2. The Panic Room (TPR)

### *2.1. Project definition*

Our main aim is to create a system based on a bottom-up "Ambient Intelligence" context that develops a proto-emotion of fear or pleasure in order to detect a dangerous event and react to it.

The project has two stages, the first of which we will develop in this chapter:

1. The first stage of the project consists in a computer simulation to check out the possibilities of such an approach before constructing a real prototype. We have made a model programmed in Python.
2. In the second stage we will create a real room, called “The Panic Room”, equipped with several components: 4 doors: in/out, 3 switches: which activate emotional responses, 4 sensors: which detect proximity to switches and movement inside the room and, finally, 4 sound and light devices: which show the emotional situation of the Room.

The main aim of the first stage of the project, the computational simulation, is to simulate - in order to be able to actually build a computational prototype - several non-conscious characteristics of emotions that are related to behaviour modification. This system is a hardwired one, that is, a circuit that is permanently connected to perform a specific function, as distinct from circuits addressed by software in a program and therefore capable of performing a variety of functions, albeit more slowly [25]. Like insects and other living beings, our system is hardwired too [26], mimicking at a basic level the possibilities of interactions with the world based upon simple emotional states.

To understand our objectives it might be worth stating first what we do not plan to do:

a) We do not plan to build a digital system which is able to "feel emotions". Qualia are still one of the "hard problems" and we would not even know where to start.
b) The system is not designed to recognise emotions in other beings.
c) It is not an expert system that is able to develop reasoning and arguments based on the concepts of emotions.

So, what are we trying to build then? Mostly we want to simulate a main characteristic of emotions: their ability to produce some behaviour aimed at adapting to a specific change in the environment. We are looking for a process that:

a) automatically evaluates a situation as being either "neutral" or "dangerous".
b) will generate an automatic response that deals with the specific danger detected.
c) can escalate the response: that is can make the response stronger or weaker depending on the degree of danger that is detected.

In this model of emotion, feeling the qualia associated with it is not needed in order to generate the response. What we are interested in is the process of emotion as a signal that is generated after an evaluation of the surroundings and is then able to produce some automatic responses. These, of course, are not all the characteristics that real emotions display, so we prefer to talk about our system as having "proto-emotions".

Another key aspect of emotions we wanted to simulate is their valence. Since Aristotle (in *Nicomachean Ethics* where he analysed the role of the passions in virtue, and in *Rhetoric* where he considered emotions as combinations of a feeling of pleasure or pain with a belief; consequently, changes in belief should change emotions), we know that emotions are either positive or negative, producing either pleasure or pain. Of course, there is more to it, as a person who has just enjoyed a roller coaster would tell you: in this case the emotion is negative in principle (fear), but it actually gives you pleasure. Again, there is a long road to travel before we can make robots that can enjoy themselves in Disneyworld, so we can safely forget that type of process from now on and develop a system that has only two emotions: one positive and one negative.

These two emotions are connected to the two main stimuli provided by the surroundings: neutral or dangerous. The more the negative emotion is felt, the closer the system is to starting to take measures to counter the danger. The system has several equilibrium levels that need to be exceeded in order to react to the possible menace. Again, these emotions are not "felt" by the system either as positive or negative. The system has no possibility of perceiving qualia. We term them negative or positive in the sense that the negative emotion forces the system to protect itself by activating a secondary electrical circuit, while the positive one tends to return it to the 'normal' status.

Therefore, when we say that our system simulates a proto-emotional circuit we mean that:

1. The system labels the signals from the sensors which describes the surroundings as either negative or positive,
2. Each option has a specific signal that is used to change the way further perceptual signals will be processed as well as to generate possible behavioural responses to a potential danger,
3. (Emotional) Responses are automatic and embedded in the system,
4. All the computations are based on the relative strength of the two proto-emotional signals. If the negative signal reaches a certain threshold it will activate the "defensive action" and will switch to the emergency circuit. Positive signals have the effect of "calming down" the system in order to avoid that reaction.

The real TPR project will use motion detector sensors and a hardwired action approach in order to interpret the movements of passers-by and decide whether they might be dangerous to the system or not. To learn and to react, the system uses a synthetic proto-emotion of fear. Fear can be considered as a preventor of danger. To certain events that in the past endangered the system, the ambient intelligence section associates a "fear signal" and the system is then directed to avoid the source of fear and react in order to avoid danger. The system uses a very basic implementation of fear. A situation that is labelled by the system as "dangerous" creates a "sensation" of fear. That sensation causes the system a) to follow the signals from the sensors a lot more

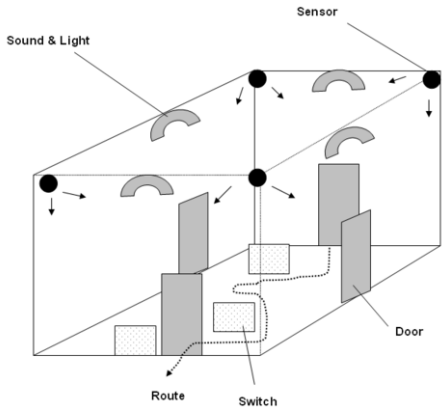



**Figure 2.** A diagram of a (future) real TPR

closely b) to increase the fear effect of any further dangerous signal and c) to consider certain signals that previously were considered harmless as potentially dangerous.

*2.2. What happens when walking through the room?*

The system simulates a room in which passers-by can walk around. There are three switches distributed around the room. If a user is able to disconnect the three switches in rapid succession then the power is cut to the main computer running the entire environmental construction and the whole system fails. However, if the system is able to detect such an attack beforehand, it has a few seconds to acquire an alternative source of electricity before the user turns off the final switch. To make the process more interesting, the system does not have access to information about whether the switches have been turned off or not.

By means of a deterministic algorithm, one not capable of change through learning we design the system to distinguish between a harmless and a harmful intruder. Each movement by the user either generates some elevation or reduction of a fear signal. As the fear increases the system checks the signals coming from the more relevant sensors more frequently. Once the signal goes beyond a certain threshold, the system enters into "panic mode" and grabs the alternative source of electricity. When the fear signal descends enough for the system to consider that the danger has passed it, returns to its normal activity, getting electricity again from the usual source.

Figure 1. shows an example of a simulated route through TPR:

1. User enters the room. He or she might have touched the first switch, in which case the system enters the first level of pain/fear.
2. User walks around the room, until he/she gets too close to the second switch. He/she might activate this switch too, causing the system to enter the second level of pain/fear.
3. User continues walking but doesn't arrive at the third switch. However with insufficient time to release the pleasure signal yet, the system continues fully activated in pain/fear.

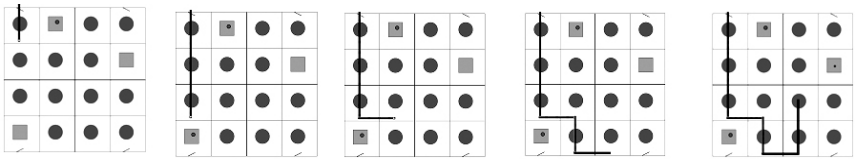


**Figure 3.** Example of a simulated route through TPR

4. User continues walking and doesn't arrive at the third switch, so the pleasure starts to overcome the pain/fear, but without eliminating it completely.
5. User gets too close to the third switch, and due to the fact that the second level signal hasn't been extinguished yet, the system enters into "panic mode" and disconnects from main energy source, turning to the emergency one.

## 3. The TPR Computational Model

Prior to the real room we have developed a computational model of the whole system. It is written with an object-oriented scripting language, Python, supported by open source software. To test the efficiency of the code, we first created a short independent program that randomly generated walks around the room. We used "Brownian noise" to produce random walks, but added several rules to avoid strange behaviour and make the walking process more human-like, avoiding journeys such as moving back and forth between one spot and another. The route started from one of the four doors in the four corners of the room and finished once the route arrived at another (or the same) door. The length of the path varied from 4 steps (a straight line from one door to another) to 16 steps, with 6 steps being the most common route.

The computational simulation is designed as a box with four doors (a,d,m,p), three switches (b,h,m), and 16 squares (a-p) (see figure 4.).

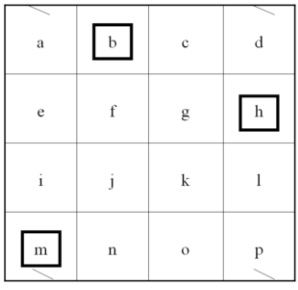



**Figure 4.**

## 4. Results

We present the results of our computational model of the TPR using a Python coded simulation of our TPR model that we ran several times in order to be able to obtain a sufficient amount of data to be studied. From more than four hundred simulations we have selected 50 different routes that make it possible to test the different emotional states and paths (see Table 2.). The data of these paths are classified in order, in/out position, the path travelled and the emotional activations of that path:

### *4.1. Data analysis*

From the data we can observe that:

- TPR is able to discriminate between neutral and pain/pleasure signals.
- Most paths are travelled in neutral states.
- TPR has a correct escalation response through [pain➔ pain+➔ panic] states.
- There is a higher tendency towards pain activations than towards boxes than pleasant ones.

**Table 2.**

| No. | In | Out | Time (sec.) | Path | Emotional states: neuter | pain | pain − | panic | pleasure | happyness |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | d | p | 30 | d,c,g,h,l,p | d,c,g,h,l,p | | | | | |
| 2 | m | m | 25 | m,i,j,n,m | m,i,j,n,m | | | | | |
| 3 | d | a | 25 | d,c,b,e,a | d | c,b,e | | | | a |
| 4 | m | p | 40 | m,i,f,j,k,h,l,p | m,i | f | j | k,h,l,n | | |
| 5 | p | a | 30 | p,o,n,I,e,a | p,o,n,i | e | | | | a |
| 6 | a | m | 20 | a,e,i,m | ae,i | | | | | m |
| 7 | d | m | 45 | d,c,b,f,j,n,o,n,d,m | d | c,b,f, | j,n,o,n | | | m |
| 8 | a | p | 35 | a,e,i,j,k,l,p | | a,e | i,j | k,l,p | | |
| 9 | p | d | 30 | p,k,f,b,c,d | p,k | f,b,c | | | | d |
| 10 | m | d | 20 | m,j,g,d | m,j | g | | | | d |
| 11 | a | p | 45 | a,b,c,d,h,g,k,l,p | | a,b,c,d,g,k,l, | | | h | p |
| 12 | p | d | 40 | p,o,k,l,h,g,c,d | p,o,k,l,h | g,c | | | | d |
| 13 | d | a | 70 | d,c,b,e,i,j,i,e,i,j,f,b,f,a | d,j | c,b,e,f,b | i,j,e | | i | a |
| 14 | p | d | 30 | p,o,k,h,g,d | p,o,k,h,g,d | | | | | |
| 15 | p | m | 20 | p,o,n,m | p,o,n,m | | | | | |
| 16 | m | a | 25 | p,k,f,b,a | p,k | f,b | | | | a |
| 17 | p | d | 35 | p,k,f,j,k,g,d | p,k | f | j | k,g,d | | |
| 18 | m | d | 45 | m,n,o,k,g,f,l,e,a | m,n,o,k,g | f,l,e | | | | a |
| 19 | p | m | 20 | p,o,n,m | p,o,n,m | | | | | |
| 20 | p | p | 20 | p,o,k,p | p,o,k,p | | | | | |
| 21 | d | d | 70 | d,c,g,k,j,i,j,f,b,c,h,d | d,h,d | c,g,k, | j,i,j,f,b | | c | |
| 22 | d | d | 20 | d,h,g,d | d,h,g,d | | | | | |
| 23 | p | d | 40 | p,o,n,j,f,g,c,d | p,o,n,j | f,g,c | | | | d |
| 24 | a | m | 20 | a,f,k,p | | a,f,k | | | | p |
| 25 | m | m | 15 | m,i,m | m,i,m | | | | | |
| 26 | a | p | 35 | a,e,i,j,k,l,h,l,p | | a,e | i,j | k,l,h,l,p | | |
| 27 | p | p | 25 | p,o,n,o,p | p,o,n,o,p | | | | | |
| 28 | p | a | 75 | p,o,k,j,n,j,k,l,h,c,b,e,i,e,a | p,o,k,j,n,j,k,l,h | c,b,e | i,e | | | a |
| 29 | d | d | 45 | d,h,l,k,g,h,g,c,d | d,h,l,k | g,h,g,c | | | | d |
| 30 | a | d | 35 | a,e,i,f,g,h,d | | a,e | i,f | g,h,d | | |
| 31 | p | d | 40 | p,k,g,f,b,c,g,d | p,k,g,d | f,b,c,g | | | | |
| 32 | a | a | 45 | a,b,c,d,h,g,f,e,a | | a,b,c,d,g,f,e | | | h | a |
| 33 | p | d | 45 | p,o,n,j,f,j,k,h,d | p,o,n,j | f | j | k,h,d | | |
| 34 | p | d | 30 | p,o,j,g,h,d | p,o,j, | g,h | | | | d |
| 35 | d | m | 50 | d,g,f,i,j,i,e,i, j,m | d | g,f | i,j,i,e,i, j | | | m |
| 36 | a | d | 20 | a,f,g,d | | a,f,g | | | | d |
| 37 | m | d | 30 | m,i,f,j,g,d | m,i | f | j | g,d | | |
| 38 | a | d | 55 | a,f,j,n,o,p,o,k,g,c,d | | a,f | j,n,o,p,o | k,g,c,d | | |
| 39 | m | m | 25 | m,i,e,j,m | m,i | e | j | | | m |
| 40 | p | m | 45 | p,o,k,g,k,l,o,n,p | p,o,k | g,k,l | | | n | p |
| 41 | m | m | 30 | m,n,o,j,m | m,n,o,j,m | | | | | |
| 42 | d | p | 30 | d,c,h,l,p | d | c,h,l | | | | p |
| 43 | m | a | 70 | m,i,e,f,j,n,j,f,b,f,e,a | m,i | e,f | j,n,j,f,b,f,e | | | a |
| 44 | m | p | 45 | m,n,o,n,o,n,j,k,p | m,n,o,n,o,n,j,k,p | | | | | |
| 45 | a | p | 40 | a,b,c,h,l,h,l,p | l,p | a,b,c,h,l | | | h | |
| 46 | a | a | 25 | a,b,e,a | | a,b,e,a | | | | |
| 47 | a | d | 30 | a,e,i,f,g,d | | a,e, | i,f | g,d | | |
| 48 | a | p | 40 | a,e,i,f,g,h,l,p | | a,e | i,f | g,h,l,p | | |
| 49 | d | d | 35 | d, c, g, k,g,c,d | d, c, g, k,g,c,d | | | | | |
| 50 | d | m | 35 | d,h,l,o,k,n,m | d,h,l,o,k,n,m | | | | | |

pleasure ones. This is the result of a square with more dangerous

- Two activation signals (pain and pleasure) are sufficient to allow the TPR to carry out a coherent *survival* activity.
- A hardwired approach to ambient intelligence is possible with TPR.

## 5. Conclusions

The system labels the signals from the sensors which describe the surroundings either as negative or positive. Each option has a specific signal that is used to change the way further perceptual signals will be processed as well to as generate possible behavioural responses to a potential danger.

Responses are automatic and embedded (or *hardwired*) in the system (they are *intentional* - but not conscious - forces). All computations are aimed at calculating the sum of the two types of proto-emotional signals (+ and -) currently influencing the system in order to establish the relevant state and trigger any appropriate physical action. If the negative signal reaches a certain threshold it will activate the "defensive action" and will switch to the emergency circuit. Positive signals have the effect of "calming down" the system in order to avoid that reaction.

Finally, we consider that this might be a first step toward the simulation of more complex emotions and the development of true synthetic emotions.

TPR can easily distinguish between dangerous and innocent situations from its basic emotional structure (using pain and pleasure). We demonstrate that emotions, as hardwired conditions of the system, are intentional maps of action that make possible an effective interaction with the world without the necessity for complex programming. We think that, in the future, it will be necessary to add the possibility of memory to the TPR, in order to increase its complexity and to investigate adaptive-learning processes through proto-emotional signals.

From these simulation data, and after the publication of this chapter, we will develop a physical device from the computational model that enables us to test it with the interaction of human beings.

## 6. Acknowledgements

This work was partly supported by TECNOCOG research group (UAB) on Cognition and Technological Environments, [HUM2005-01552]. Thanks for the very interesting questions and suggestions made through the online comments of our computer and engineering students (Philosophy and Computing, UAB) and to Peter Skuce.

## References

[1] A. Damasio, *Descartes error,* MIT Press, Cambridge (MA), 1994.
[2] G. Edelman, G. Tononi, *A Universe of Consciousness: How Matter Becomes Imagination,* Basic Books, New York, 2000.
[3] D. Denton, *Emotions: The Dawning of Consciousness*, OUP, Oxford, 2006
[4] V.S. Ramachandran, *A Brief Tour of Human Consciousness*, Pi Press, Pearson Education, New York, 2004.
[5] R, Picard, *Affective Computing*, MIT Press, Cambridge (MA), 1997.

[6] C. Breazeal, *Designing sociable robots*, MIT Press, Cambridge (MA), 2002.
[7] A. Sloman, Interactions between Philosophy and AI, Artificial Intelligence, 2 (1971), 209-225.
[8] A, Adamatzky, A., 2003, Affectons: automata models of emotional interactions, Applied Mathematics and Computation, 146 (2003), 579-594.
[9] Elliot, C., 1992,, *The affective reasoner: a process model of emotions in a multi-agent system,* PhD Thesis Northwestern University, Mayo, The Institute for Learning Sciences, Technical report #32, 1992.
[10] M. Scheutz, A. Sloman, Affect and agent control: Experiments with simple affective states, in N. Zhong, J. Liu, O. Ohsuga, J. Bradshaw, J. (eds.), *Intelligent agent technology:Research and development*, World Scientific Publisher, NJ, 2001, 200-209
[11] S. Walczak*, "Modeling Affect: The Next Step in Intelligent Computer Evolution",* Informatica*,* 9: 4 (1995), 573-584
[12] J-M. Fellous, M.A. Arbib, *Who Needs Emotions? The Brain Meets the Robot,* OUP, Oxford, 2005.
[13] M.D. McNeese, New visions of human-computer interaction: making affect compute, *Int. J. Human-Computer Studies*, 59 (2003), 33-53.
[14] K. Oatley, The bug in the salad: the uses of emotion in computer interfaces, *Interacting with computers*, 16 (2004), 693-696.
[15] R.A. Brooks, Intelligence without representation, *Artificial Intelligence*, 47 (1991), 139-159.
[16] A. Sloman, M. Croucher, Why robots will have emotions, *Proceedings 7th International Joint Conference on AI,* Morgan-Kaufman, USA, 1981.
[17] C. DeLancey, *Passionate Engines: What Emotions Reveal About Mind and Artificial Intelligence,* OUP, Oxford, 2001.
[18] A. Sloman, R. Chrisley, M. Scheutz, Architectural basis of affective states & processes, paper for inclusion in Fellous and Arbib (eds), *Who Needs Emotions?: The Brain Meets the Machine*, OUP, Oxford, 2003.
[19] H.L. Dreyfus, *What Computers Still Can't Do*, MIT Press, Cambridge (MA), 1994.
[20] J. McCarthy, J. Ascribing mental qualities to machines, in M. Ringle (ed), *Philosophical Perspectives in Artificial Intelligence*, Harvester Press, USA,1979
[21] J. McCarthy, P. Hayes, Some philosophical problems from the standpoint of artificial intelligence". In B. Meltzer & Michie (eds.), *Machine Intelligence*, vol. 4, Edinburgh University Press, Edinburgh, American Elsevier, New York, 1969, 463—502
[22] G. Rizzolati, L. Craighero, The Mirror-Neuron System, *Annual Review of Neuroscience*, 27 (2004), 169–192
[23] P.Thagard, The passionate scientist: emotion in scientific cognition, In P. Carruthers, S. Stich, M. Siegal (eds.), *The Cognitive Basis of Science*, Cambridge University Press, Cambridge, 2002.
[24] P. Thagard, F.W. Kroon, J. Nerb, *Hot Thought: Mechanisms And Applications of Emotional Cognition*, MIT Press, Cambridge (MA), 2006
[25] *Dictionary of computing*, OUP, Oxford, 1986
[26] W.C. Clark, M. Grunstein, *Are We Hardwired? The Role of Genes in Human Behavior*, OUP, Oxford, 2000
[27] A. Damasio, *The Feeling of what happens: Body and Emotion in the making of Consciousness,* Harvest, San Diego, 1999
[28] K. Oatley and J. Jenkins, *Understanding emotions,* Blackwell, Cambridge (MA). 1996
[29] F. Sousa, *The Rationality of emotions,* MIT Press, Cambridge (MA), 1987

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*


# Representation in Digital Systems

Vincent C. MÜLLER[1]
*American College of Thessaloniki*

**Abstract.** Cognition is commonly taken to be computational manipulation of representations. These representations are assumed to be digital, but it is not usually specified what that means and what relevance it has for the theory. I propose a specification for being a digital state in a digital system, especially a digital computational system. The specification shows that identification of digital states requires functional directedness, either for someone or for the system of which the state is a part. In the case of digital representations, the function of the type is to represent, that of the token just to be a token of that representational type.

## 1. Digital Representations and the Computationalist Program

In this paper, I will attempt to clarify one aspect of a notion that is commonly used in the cognitive sciences, but has come under considerable criticism in recent years: the notion of representation. Representations are typically invoked in a 'computational theory of the mind', where mental processes are understood as information processing through computational operations over *representations.* The standard theory is the computational representational theory of mind (CRM or "computationalism"), which says that the human mind is a functional computational mechanism operating over representations. These representational abilities are then to be explained naturalistically, either as a result of information-theoretical processes [1,2], or as the result of biological function in a "teleosemantics" [3,4].

While there is intense discussion about what role representations play and whether their manipulation must be computational, there is one aspect that is commonly glossed over: The computational processes in question are taken to be *digital* computational processes, operating algorithms over *digital* representations. But what constitutes a digital computation and a digital representation? And if this were specified, does the specification have repercussions for the computational representational theory of the mind?

Such repercussions are to be expected in several areas that are crucial for a computational theory of the mind. One of these is the question whether something can be called a digital state at all without *presupposing* mental processes – in which case there is a threat of a circle. Another is the problem of "grounding". This alleged problem is intimately connected with the current wave of 'embodied cognition'. A concise formulation probably still is: "How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?" [5]. Now, we have argued in recent papers on nonconceptual phenomenal content [6,7] that a non-conceptual content should be at the base of such grounding: Such content is retrieved bottom-up (to "cognitively encapsulated" modules), it is not mediated top-down by concepts, is independent of conceptual resources available to the person, and not phenomenal (accessible to the person). The nonconceptual representations of objects can provide a starting point for the grounding procedure that does not assume conceptual material (but rather things like spatiotemporal information of objects such as existence, persistence in time and through motion, spatial relations relative to other objects, movement; basic information on surface properties, shape, size and orientation). If it were to come out that such content is necessary but cannot be present in purely digital systems, this would show that human cognition is not purely digital – and that artificial intelligence on purely digital computers is impossible.

[1] Corresponding Author: Vincent C. Müller: American College of Thessaloniki P.O. Box 21021, 55510 Pylaia, Greece, http://www.typos.de, E-mail: vmueller@act.edu

We can take our starting point from the recent debate whether mental representations are material [8,9], or whether thinking perhaps proceeds via conventional linguistic symbols rather than especially 'mental' symbols on a sub-personal level [10]. Sedivy, in particular, argues that cognition does not proceed as a manipulation of 'carriers' of representation: In the case of linguistic symbols, for example, there is a physical object (ink on a page, compressions of air, etc.) that carries content, but in the case of mental representation at the personal level there is not. Could digital states be such carriers?

The question of digital representation is also relevant for the general question which physical objects in the world are computers; a problem that Shagrir calls "problem of physical computation" [11]. Digital computation necessarily involves digital representation on the classical view: "There is no computation without representation" [12]. O'Brien even defines computing with the help of representation: Computation is "a procedure in which representational vehicles are *processed in a semantically coherent fashion*" [13] (a "coherent fashion" being one in which they "bear comprehensible [non-arbitrary] semantic relations to one another"). This view is opposed to the ones that dispense with representation, either understanding computation as a purely syntactic procedure, or widening the notion to the extent that "every natural process is computation in a computing universe" [14]; a view that is now called "pancomputationalism".

## 2. Digital Representations and Digital Content

If a digital state is part of a system in which it plays a representational role, then it will have content. This content, in turn, is necessarily digital as well since a digital state can only represent whether something is *of a type*. This applies whether or not the system is a cognitive system or not. (The red warning light on the dashboard that informs me of the engine overheating is part of a system, indeed a computational information processing system, but presumably not of a cognitive one.) The characteristic feature of a digital representation in a system is thus that it 'checks' whether an informational input is of a type or not, it *categorizes* the input. So, a digital representation can only represent a digital content; an analog content can only be approximated. For example, it is impossible to describe completely in words what a picture depicts; as the saying goes "a picture is worth a thousand words", i.e. an analog representation can be categorized in a thousand ways, and yet none will be sufficient. Any indication to the contrary is produced by that fact that one cannot *say* with words what is the analog content that is not represented; one can just *indicate* that there is such content by pointing out what is missing, in each representation.

Dietrich and Markman summarize their discussion with the statement "If a system *categorizes* environmental inputs, then it has discrete [digital] representations" and "A system has discrete representations if and only if it can discriminate its inputs." [15] This is appropriate, but it *presupposes* that the digital states in question are representations. What we need to find out is whether this is accidental, or if *all* digital states are representations?

The notions of representation and of representational content are, of course, disputed, but it seems basic to distinguish (in the tradition of C. S. Peirce) *symbolic* and *iconic* representation from *indices* (e.g. smoke indicates the existence of fire). I would propose to capture this difference by calling the former representation and the latter information.

On this terminology, information is whatever can be learned from the causal history of an event or object, representation is what it is *meant to represent* (in a suitable notion of function). It is clear that in this usage, information is always true, while a representation may not be. (So, if someone lies to you, he is not giving you false information, he is misrepresenting the world.)

Nonetheless, we can still identify (at least) two notions of representation. One typical version of a wider notion is: "a representation is any internal state that mediates or plays a mediating role between the system's inputs and outputs in virtue of that state's [explicit] semantic content" [15]. On the other hand, there is a use of the term in which symbols and (more or less interpretation-involving) icons are representations for a person. As I said in a different context:

> "…we need to distinguish between a representation *per se*, and a representation *for someone*. If we are wondering whether, say, a squiggle on a piece of paper is a representation, this is (normally) to ask whether it is a representation for someone. If we are wondering whether a set of switches or a neural pattern represents, this is to ask whether it has the function to represent in a larger system." [16]

Note that, in the end, both notions involve functional talk: being a representation for someone and being a representation for the system. I will argue presently that this is characteristic of digital states.

There is a common view that couples representation quite generally with digital processing, e.g. "talk of representations [in the analysis of cognitive functioning] invites speculation about content and reference which leads to a symbolic model of cognitive processing, or at least blurs the distinction between analog and symbolic simulation," [17]. As we shall see shortly, even if there are digital representations, there are also others, probably even in cognitive systems.

### *2.1. Discreteness vs. continuity*

In a first approximation, being digital means being in a discrete state, a state that is strictly separated from another, not on a continuum. Prime examples of digital representations are the states of a digital speedometer or watch (with numbers as opposed to an analog hand moving over a dial), the digital states in a conventional computer, the states of a warning light, or the states in a game of chess. Some digital states are binary, they have only two possible states, but some have many more discrete states, such as the 10 numbers of a digital counter or the 26 letters of the standard English alphabet.

It is characteristic of such digital representations that they can have multiple realizations. So, one can write the same word twice, even though one cannot make exactly the same mark on paper twice. The possibility of multiple realization is a *result* of digital states being discrete: Since a mark on a piece of paper can be clearly a "T", we can rub it out and replace it with a new "T", or copy the "T" to a new piece of paper. It does not matter that there are borderline cases which are neither clearly this letter nor clearly another, as long as there are clear cases. (This becomes easier if one knows *a priori* that something is meant to be a letter, i.e. a digital representation from a particular set.)

Crucially, a digital representation is a *token of a type*. Being such a token or not is what allows it to be discrete. Being such a token allows it to be realized in multiple ways and several times. *Which* types exist is often pre-defined, as in the case of the 26 letters of an alphabet. Something either constitutes a token of one of these types or it is not a letter of this alphabet at all.

### *2.2. Analog representations*

Digital representations are, to characterize them negatively, not analog representations. But what does that mean? Trenholme says that "... analog simulators are physical systems whose processes are characterized by what might be termed *variable properties* - physical magnitudes whose values vary over time." [17], meaning that they vary over time in a non-discrete way, *and* in analogy to the represented. For example, the movement of the hand over the speedometer is analogous to the speed of the vehicle – while the digital numbers on a display are not in such an analogy.

Fodor points out that the analogy is a product of physical laws, and that it takes the absence of such laws to indicate a digital representation: "A machine is analog if its input-output behavior instantiates a physical law and is digital otherwise" [18]. Another formulation is that an analog process is "one whose behavior must be characterized in terms of lawful relations among properties of a particular physical instantiation of a process, rather than in terms of rules or representations (or algorithms)." [19] This characterization is used by Demopoulos who defines a digital machine class (a class of machines with the same digital states) by saying that their "behavior is capturable only in computational terms" [20], i.e. not by physical laws.

These remarks are very plausible, if we consider that in the case of a continuous representation (e.g. the analog speedometer), the relation between what is represented and what represents is wholly determined by the physical laws governing the two and their "correspondence" [15]. It is only when digital states come into play that the continuous input is 'chunked' into types. This division into tokens of types in a digital system excludes its description purely in terms of physical laws precisely because it also requires a reference to function or directedness (it is not accidental that the rejection of psychophysical "identity theories" was the starting point of functionalism in the philosophy of mind).

But which of the two characteristics is crucial for an analog state, the analogy to the represented, or the continuous movement?

This question becomes relevant in the case of representations that proceed in "steps" but also in analogy to the represented, e.g. a clock the hands of which jump from one discrete state to another. Zenon Pylyshyn argues that the underlying process is analog, and this is what matters: "an analog watch does not cease to be analog even if its hands move in discrete steps" [21,22]. James Blachowicz also thinks that being on a continuum is sufficient for being analog, taking the view that "differentiated representations may also be analog – as long as they remain *serial*", his example is a slide rule with "clicks" for positions [23].

Note, however, that the very same underlying mechanism could give a signal to a hand to move one step and to a digit to go one up (this is actually how clocks are controlled in centralized systems, e. g. at railway stations). Both of these would be analogous to the "flow of time", given that the natural numbers of the digital display are also in a series. So, while the underlying mechanism might be analog or digital, the question whether the representation is digital ultimately hinges on whether it is discrete, not on whether it is analogous to the represented. From what we have seen thus far, a digital representation is just a discrete representation.

### *2.3. Digital and analog - on which level?*

Returning to the classical computational representational theory of the mind, it is important for our analysis to see that a computer can be described on several levels and identify which one is relevant for us here.

First of all, a computer can be described on the *physical level:* Some physical objects such as toothed wheels, holes in cards, states of switches, states of transistors, states of neurons, etc. are causally connected with each other such that a state of one object can alter the state of another. So far, this is just any system. At the s*yntactical level,* the states or physical objects are taken to be tokens of a type (e.g. charge/no charge) and are manipulated according to algorithms. At first glance, this manipulation only concerns these tokens; it is "purely syntactical." Finally, there is a s*ymbolic level* of what the objects and states that are manipulated on the syntactical level are taken to represent, e.g. objects in the world. Perhaps it would be more appropriate to say that there are several symbolic levels, since one symbol can represent another, that can represent, say, a color, that represents, in turn, a political party, etc.

It so happens, for technical reasons, that the types that we humans most frequently use to represent (e.g. natural numbers, letters, words) are normally not the types that are syntactically manipulated in a conventional computer, but are rather represented, in turn, in a further system of types, a "binary system" with tokens of just two types (on/off, 1/0, etc.). At *both* these levels, we have digital representation. [In the slogan "There are exactly 10 kinds of people in the world, those that understand binary and those that do not", the "10" is a binary sequence of two bits, representing the number two.]

## 3. Digital States as Tokens of a Functional Type

It is useful to note that not all systems that have digital states are digital *systems*. We can, for example, consider the male and female humans entering and leaving a building as digital states, even as a binary input and output, but in a typical building these humans do not constitute a digital system because a relevant causal interaction is missing. In the typical digital system, there will thus be a *digital mechanism*, i.e. a causal system with a *purpose*, with parts that have *functions*. Digital mechanisms in this sense may be artifacts (computing machines) or natural objects (perhaps the human nervous system). However, even if all digital representations are part of digital systems, they are not all part of computational systems – the letters and words on this page are an example in point.

We need the notion of a system because the notion of "being of a type" is too broad for our purposes. If being of a type were the criterion for being digital, then everything would be in any number of digital states, depending on how it is described. However, what we really should say is that something is digital because that is its particular *function*. The first letter of this sentence is in the digital state of being a "T" because that is its function – as opposed to an accidental orientation of ink or black pixels (or Putnam's ants making apparently meaningful traces in the sand [24]).

Some of the systems are artifacts, made for our purposes, where some physical states cause other physical states such that these function as physical states of the same set of types (e.g. 1 or 0). (Note that one machine might produce binary states in several different physical ways, e.g. as voltage levels and as magnetic fields.) If someone would fail to recognize that my laptop computer has binary digital states, they would have failed to recognize the proper (non-accidental) function of these states for the purpose of the whole system – namely what it was made for.

So, the function is what determines whether something is a token of a type or not. The normativity of having or fulfilling a function generates the normativity of being of a type. The type has the function; being of the type allows fulfilling the function.

### *3.1. Which function?*

In the case of an artifact, we *assume* a functional description. If the engine warning light on a car dashboard is off, is it in a digital state? Yes, if its function is to indicate that nothing is wrong with the engine temperature. (It may serve all sorts of other accidental functions for certain people, of course.) But if the light has no electricity (the ignition is off), or if it was put there as a decorative item, then the lamp is not in a digital state "off". It would still be off, but this state would not be digital, would not be a token of the same functional kind.

The description of an artifact in terms of function is to say that something is a means to an end; it serves the function to achieve that end – a function that can be served more or less well. (Note that serving a function does not mean being used for that function; there may well be no agent that can properly be said to be using the artifact, e.g. if it is part of a large and complex system.) In the case of a natural object, the allocation of proper function (as opposed to "accidental function") is dependent on teleological and normative description of systems [cf. 25] – a problematic but commonplace notion.

The interesting cases for our understanding of digital states are those where it is not clear that some object or state really is a digital representation. The prehistoric cave paintings contain repeated patterns, so are these symbols? Are what seem to be icons actually digital representations, as it turned out in the case of the Egyptian hieroglyphs? And what about the neural activity in the brains of humans and other mammals? The function of a human's brain seems to be cognition, so whether the brain is in digital states depends on whether its cognitional function is fulfilled in this way. Are there *types* of neural states that serve a representational function? (Piccinini has recently presented interesting evidence that this is not the case [26].)

### *3.2. Representational function*

The problem of identifying a function could perhaps be solved if we helped ourselves to representation as the function. (I am grateful to Philip Brey for a question in this direction at E-CAP 2007.) Now, I am *not* of the view that we should restrict the notion of digital states only to the representational ones because I think that digital states in several characteristic areas, in particular in conventional digital computers are *not* representational.

Nonetheless, in the cases where the digital states *do* serve representational function, it will be precisely *that* function that determines being or not being of a type. So, as I said above, when we wonder whether something is a letter, the answer depends on whether it serves a representational function as a token of a particular (finite) set of types. In that case, just being of the type is to fulfill the function. So, if a state has that function *qua* token of a representational type, then it is a digital representation.

This understanding should not be taken as a motivation to re-introduce semantical notions into the definition of digital states or computing. It has been argued in many places that a computational mechanism must involve semantics or meaningful symbols, because this is necessary for its supposed carrying out of orders, or for its identification of tokens (e.g.[27,28] and [29,30]). Kuczynski even argues that there is really no such thing as a purely formal procedure because semantics is needed to identify tokens, which leads him to say, in the end, that: "Since the Turing machine is not semantics-driven, it is not responding to logical form. Therefore it is not computing (in the relevant sense), even though, from some viewpoint, it acts like something that is computing" [31].

These consequences do not follow if we recall that digital states are situated at the level of syntactic description. The level of syntactic description can be identified with the help of representational function without thereby saying that there is somehow a semantic or representational level in the machine. The Turing machine computes, but its computations mean nothing to it; it is only to us that they are, for example, operations over natural numbers.

**Acknowledgements**

I am grateful to audiences at the universities of Tübingen and Mälardalen as well as at the E-CAP (Twente) and NA-CAP (Chicago) conferences for useful comments, especially to Alex Byrne and Kurt Wallnau. My thanks to Gordana Dodig-Crnkovic and Luciano Floridi for written comments on a related paper. Thanks to Bill Demopoulos also. I am very grateful to the discussions on a related paper that was presented at the "Adaptation and Representation" web conference [32], especially to Gualtiero Piccinini.

**References**


[1] F. Dretske, *Knowledge and the flow of information*, MIT Press, Cambridge, Mass., 1981.
[2] F. Dretske, *Naturalizing the mind*, MIT Press, Cambridge, Mass., 1995.
[3] R.G. Millikan, *Language: A biological model*, Oxford University Press, Oxford, 2005.
[4] G. Macdonald, D. Papineau, editors, *Teleosemantics: New philosophical essays*. Oxford University Pres, Oxford, 2006.
[5] S. Harnad, The symbol grounding problem, *Physica D* 42 (1990), 335-346.
[6] A. Raftopoulos, V.C. Müller, The phenomenal content of experience, *Mind and Language* 21 (2006), 187-219.
[7] A. Raftopoulos, V.C. Müller, Nonconceptual demonstrative reference, *Philosophy and Phenomenological Research* 72 (2006), 251-285.
[8] A. Clark, Material symbols, *Philosophical Psychology* 19 (2006), 291-307.
[9] S. Sedivy, Minds: Contents without vehicles, *Philosophical Psychology* 17 (2004), 149-179.
[10] J. Speaks, Is mental content prior to linguistic meaning?, *Nous* 40 (2006), 428-467.
[11] O. Shagrir, Why we view the brain as a computer, *Synthese* 153 (2006), 393-416.
[12] J.A. Fodor, The mind-body problem, *Scientific American* 244 (1981), 114-123.
[13] G. O'Brien, Connectionism, analogicity and mental content, *Acta Analytica* 22 (1998), 111-131.
[14] G. Dodig-Crnkovic, Epistemology naturalized: The info-computationalist approach, *APA Newsletter on Philosophy and Computers* 6 (2007), 9-14.
[15] E. Dietrich, A.B. Markman, Discrete thoughts: Why cognition must use discrete representations, *Mind and Language* 18 (2003), 95-119.
[16] V.C. Müller, Is there a future for AI without representation?, *Minds and Machines* 17 (2007), 101-115.
[17] R. Trenholme, Analog simulation, *Philosophy of Science* 61 (1994), 115-131.
[18] J.A. Fodor, N. Block, Cognitivism and the digital/analog distinction, unpublished, cited in [20].
[19] J.A. Fodor, Z. Pylyshyn, How direct is visual perception?, *Cognition* IX (1981), 139-196.
[20] W. Demopoulos, On some fundamental distinctions of computationalism, *Synthese* 70 (1987), 79-96.
[21] Z.W. Pylyshyn, *Computation and cognition*, MIT Press, Cambridge, Mass., 1984.
[22] O. Shagrir, Two dogmas of computationalism, *Minds and Machines* **7** (1997), 321-344.
[23] J. Blachowicz, Analog representation beyond mental imagery, *The Journal of Philosophy* 94 (1997), 55-84.
[24] H. Putnam, *Reason, truth and history*, Cambridge University Press, Cambridge, 1981.
[25] U. Krohs, Der Funktionsbegriff in der Biologie, in: A. Bartels, M. Stöckler, editors, *Wissenschaftstheorie: Texte zur Einführung*. Mentis, Paderborn, 2007, forthcoming.
[26] G. Piccinini. Digits, strings, and spikes: Empirical evidence against computationalism. NACAP Conference. Chicago; 2007.
[27] M.A. Boden, Escaping from the Chinese room, in: M.A. Boden, editor, *The philosophy of artificial intelligence*. Oxford University Press, Oxford, 1990, 89-104.
[28] M.A. Boden, *Mind as machine: A history of cognitive science*, Oxford University Press, Oxford, 2006.
[29] J. Haugeland, *Artificial intelligence: The very idea*, MIT Press, Cambridge, Mass., 1985.
[30] J. Haugeland, Syntax, semantics, physics, in: J. Preston, M. Bishop, editors, *Views into the Chinese room: New essays on Searle and artificial intelligence*. Oxford University Press, Oxford, 2002, 379-392.
[31] J.-M. Kuczynski, Two concepts of 'form' and the so-called computational theory of mind, *Philosophical Psychology* 19 (2006), 795-821.
[32] V.C. Müller, 2007, *Representation in digital systems (with comments and replies),* in Interdisciplines: Adaptation and representation, Institut des Sciences Cognitives/CNRS, Université de Genève <http://www.interdisciplines.org/adaptation/ papers/7>, accessed 11.07.2007.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*


# Information, Knowledge and Confirmation Holism[1]

**Steve MCKINLAY**
*School of Information Technology, Wellington Institute of Technology, Petone, New Zealand*
*Charles Sturt University*
*Centre for Applied Philosophy and Public Ethics*
Email: steve.mckinlay@weltec.ac.nz

**Abstract**: An emerging alternative to the problem of knowledge looks towards *information* as playing a critical role in support of an externalist epistemology, a new theory of knowledge that need not rely upon the traditional but problematic tenets of belief and justification. In support of this information-theoretic epistemology, the relationship between information and knowledge both within philosophical and information technology scholarship has been viewed as an asymmetric one. This relationship is captured by the commonsense view that objective semantic information is prior to and encapsulated by knowledge [1]. This paper develops an argument that challenges this asymmetric assumption. Drawing on the ideas of Gareth Evans [6] and Timothy Williamson [5] we shall argue that (at least in some cases) a coextensive relationship must exist between information and knowledge. We conclude with the view that this relationship throws up problems similar to those discussed by Quine [15] in relation to confirmation holism.

**Keywords.** Semantic information, information continuum, Evans, Williamson, Floridi

## 1.

It may seem intuitive and self-evident to many that to have knowledge we *first* must have information. Although few in philosophy have dabbled with a detailed analysis of the connections between information and knowledge the intuitive account at least seems to suggest a linear, asymmetrical, and somewhat hierarchical, yet thus far largely unexplained relationship structure existing between *data*, *information* and *knowledge*, what we shall call, following Floridi [1] *the information continuum.* We immediately note that although a 'continuum' *per se* need not explicitly exhibit a hierarchy or bias in any particular direction in this particular case a clear progression is imagined from

[1] I thank my supervisors Steve Clarke and John Weckert for advice and comment regarding this paper, the CAPPE Institute for helping fund my research, my employer Wellington Institute of Technology for providing valuable time and the two anonymous referees for thought provoking and helpful comments.

*data,* perceived as having nominal epistemic value through to *knowledge* recognised as a much rarer and more valuable epistemic commodity.

Most would certainly agree with the commonsense notion that knowledge is something more than information. Furthermore the view is largely uncontroversial and widespread within our technology driven information based society. Such fashionable assumptions however often offer up a veritable philosophical orchard, ripe for the picking. Whilst the *information continuum*, as a kind of handy model may well provide a practical and convenient scaffold upon which we attempt to build our knowledge and information based systems it does not detract us from a genuine philosophical worry about the precise nature of the relationship between its three constituents. In this paper we primarily focus on the right hand side of the continuum, the side that assumes an apparent asymmetry whereby information is alleged to precede knowledge. The big picture question worthy of consideration and one that we shall keep in mind (this question is first considered in Floridi's [1] Problem 13) is can we formulate an *information first* approach to epistemology, or in other words can (or indeed should) epistemology presuppose a theory of information? Such an account would be on the face of it, external or objective in nature and thereby not reliant upon any analysis of belief or justification, concepts that have proven to be historically problematic with respect to epistemology.

On the other hand what if it turns out to be the case that the asymmetry is wrong, that information does not always precede knowledge, that instead knowledge at times takes on a much more primordial nature. Clearly there would be philosophical implications, for epistemology, information ethics and the philosophy of information. Furthermore there are likely be significant implications for information technology, which bases much of its practice and scholarly effort upon the assumption. The aim of this paper is precisely to investigate this possibility. In short the paper will present an argument that the relationship between information and knowledge is in fact coextensive. The corollary to this is that the assumption of any strict asymmetry between information and knowledge is at best a convenient myth.

Perhaps some readers even if only having had a brief acquaintance with formal philosophy have heard epistemology's old slogan. To the question 'What is knowledge?' the temptation may be to answer, 'Why justified, true belief of course'. It is indeed true that epistemologists have historically looked toward belief, justification and truth in their attempts to characterise knowledge and although this originally Platonic conception of knowledge still holds sway with many thinkers, there has been a recent exodus from the position.

The first significant challenge to epistemological orthodoxy came in 1963 with a surprisingly brief article entitled "Is Justified True Belief Knowledge?" [2]. Consider the popularly stated tripartite schema *S* knows that *P* if and only if:

1. P is true
2. S believes that P, and
3. S is justified in believing that P.

Gettier's article provided two counterexamples to the schema in order to show how it could not constitute *sufficient* conditions for the truth of the proposition, *S* knows that *P*. Gettier's examples sought to demonstrate the fallibility of justification, that is, if *S* could plausibly be justified in believing *P* but *P* turned out to be merely accidentally true, then any such construal no matter how *prima-facie* justifiable could not constitute

knowledge. According to Dretske [3], "the truth of what one believes may be quite unrelated to one's grounds (justification) for believing it."

It seemed we needed a new conception of knowledge. One such approach is to take an externalist view of the conditions required to support knowledge claims. The externalist would claim that the required conditions exist outside the bounds of the psychological states of the knowledge claimer (psychological states such as holding a particular belief). And so the externalist might frame an account of knowledge as follows: *K* knows that *s* is *F* is equal to *K*'s belief that *s* is *F* in that it is caused in some special kind of way by the relevant external facts [3,4]. These relevant external facts, or casual factors must exist external to the machinations of the mind of a knower and so constitute appropriate belief-independent knowledge-yielding conditions required to avoid the Gettier problem.

Knowledge on this account is therefore equated in some particular way with special relevant causal factors in the sense that the knowledge itself is what *justifies the belief*, more so than the sense in which it *gets justified*. Beliefs on such an account could be justified absolutely if and only if they are justified by prior knowledge. On the externalist account knowledge is reliant upon the genuine reliability of all relevant objective supporting evidence (we might propose 'information' as fulfilling this role) and not upon any deeper metaphysical analysis of belief, truth or justification.

As our starting point we take as given the *general definition of information* (GDI hereafter), elaborated by Floridi [1] as "…namely, the view that semantic information can be satisfactorily analyzed in terms of well-formed, meaningful, and truthful data. This semantic approach is simple and powerful enough for the task at hand." There are other accounts of information; Shannon's mathematical theory of communication (MTC) for example considers information as an encoded physical entity. Its main concern is with the transmission and receiving of encoded data. The MTC however does not address issues relating to meaning and reference or how semantic information can be related to truth and knowledge. Since these are the issues most central to this essay we shall leave the MTC aside.

In section 2 we briefly discuss issues related to the *data* → *information* side of the continuum as well as introducing 'Russell's Principle', what Gareth Evans terms *discriminating knowledge*: the capacity we have to pick out individual objects against background conditions. We then shift our focus to the relationship between information and knowledge. We consider the possibility that information might play a role as objective 'evidence' required for a non-doxastic (non-belief based), externalist account of knowledge.

Contraposing an *information-first* approach to epistemology is Timothy Williamson's [5] *knowledge-first* approach. Williamson defends a principle whereby one's *evidence* (what in fact looks very similar to a claim that one's relevant *information*) is equal to one's knowledge. If Williamson is correct, then knowledge, so it would seem, is in fact either a prior requirement such that the knower could claim to be in an *informed* state, or if evidence is, as I suggest, just the same as information then on Williamson's account knowledge and information are somehow *coextensive*. We consider what this means in Section 3. Whatever the answer it surely cannot be the case that both the information-first and the knowledge-first approach to epistemology are true.

In not too dissimilar fashion Gareth Evans [6] suggests that in order for an informee to understand any information-yielding event, be it a proposition or the perception of an object, photograph or whatever, *some prior information* must already

be held by the informee. Evans [6] argues the informee evaluates and appreciates the remark (in the case of a proposition[2]) according to the content of the relevant information and its relationship with information already in his possession.

Williamson [5] reverses conventional epistemology by arguing that knowing is a *sui generis* mental state, meaning that 'to know' cannot be explained with appeals to any number or combinations of internal psychological states that the knower might be experiencing for example belief or belief combined with justification, nor in combination with external (for example environmental) conditions. Instead Williamson treats knowledge as prime and he uses this position to develop his accounts of evidence and justification. Williamson however was not the first to suggest the belief / knowledge relationship as conceived by the traditionalists was the wrong way around. The operations of the informational system according to Evans [6] are far more primitive, "It is as well to reserve 'belief' for the notion of a far more sophisticated cognitive state: one that is connected with (and, in my opinion, defined in terms of) the notion of *judgement*, and so, also, connected with the notion of *reasons"*.

Thus we see in Williamson and Evans an analysis of knowledge that breaks with conventional epistemology whereby belief was traditionally assumed prior to knowledge. Evans on the one hand derives his arguments from a detailed consideration of the role reference plays in knowledge and information. Williamson dispenses with the notion of knowledge analysed in terms of justified true belief in favour of the thesis whereby *knowing* is understood as an irreducible mental state equated to evidence. Both are non-doxastic, externally oriented approaches, however both need reconciling with the current philosophical notion of semantic information. This paper draws on the ideas of Evans and Williamson in respect of the apparent asymmetry in the information continuum and concludes with a consideration of confirmation holism, which seems to emerge as a result of the discussions relating to Russell's Principle and the coextensiveness of information and knowledge.

## 2.

Quine [7] makes the distinction between the so-called *theory of reference* and the *theory of meaning*. Central concepts belonging to the later include meaning itself, synonymy, significance, analyticity and, entailment. To the prior belong the concepts of truth, denotation and extension. Note here however that Quine makes no mention of the concepts belief or justification typically associated with epistemology, in fact the role of justification is notably absent in Quine's epistemology, knowledge for Quine was clearly a matter of semantics.

Quine goes on to suggest, "given any two fields, it is conceivable that a concept might be compounded of concepts from both fields", and although Quine reckons the potential hybrid concept to a theory of meaning for Gareth Evans the concept of

[2] There of course are other arguments as to how information might be represented, arguments for example that claim evidence or information is not always necessarily propositional in nature, those discussions are beyond the scope of this article. For the time being I am assuming that all evidence, and all information can be adequately represented propositionally, this interpretation of information squares both with the GDI as well as with our popular computational representations of information, for example information stored in databases.

*information* emerges out of this murky conjunction we often call semantics. For Evans information plays a role ultimately critical within his theory of reference[3].

Nevertheless Quine [7] seems to gaze in a different direction in respect to theories of reference and meaning, the problems related to ontological commitment loom large - any given existential statements presuppose objects of a given kind, that is the sentence, which follows a quantifier, is true of some objects of that kind[4]. It is our goal here, however, to examine the apparent asymmetry associated with the information continuum and so following Evans [6] we draw an important line in the sand: we shall ignore questions of ontology.

> I have supposed myself to be working within a scheme of interpretation for the language which fixes the interpretation of, and hence fixes the objects capable of satisfying, its predicates; the questions which I want to discuss arise after these decisions have been made. [6]

Although Evans's decision to ignore questions of ontology seems somewhat arbitrary we find support for putting ontological issues aside in Floridi's [8] principle of *ontological neutrality.* Under the GDI the definition of a datum is cast as $d = (x \neq y)$ where $x$ and $y$ are two uninterpreted variables representing a difference. A datum thus is a *relational entity* however the definition leaves underdetermined the kinds of support required for the implementation of the inequality. That is, a datum is not bound by any ontological commitment. Floridi garners support for this position from a variety of thinkers, perhaps most eloquently phrased by Wiener (in Floridi's article) [8] "Information is information, not matter or energy. No materialism which does not admit this can survive at the present day."

Consider for a moment the notion of a datum as a relational entity. Inherent in such a notion is the basic assumption that it is impossible for one to make a judgement about any *thing* at all (a datum, entity, object or whatever) unless he *knows which* thing it is that the judgement is about. A relational entity is a distinguishable object, first and foremost distinguishable from other entities. Following Floridi [8] that $x$ is indeed not equal to $y$ represents a lack of uniformity between two signs or signals. Thus in order to pick some thing out as a datum *per se* one must be in a position to disambiguate that datum from surrounding background conditions, which in turn constitute data.

Evans [6] noted as much and called this "Russell's Principle" and there are some basic interpretations to which we might appeal. Firstly, our commonsense usage of the term '*knows which*' does not automatically assume to presuppose that the individual does in fact know which item he is making a judgement about – it maybe that the individual merely perceives a difference. Alternatively information may originate from an object $x$ but the content may not refer to the object particularly well just in case some error or problem occurs either at the source or with the information transmission.

On the other hand we could simply argue that none of this matters. If indeed there is some thing $x$ that the individual has observed and is consequently considering, then

---

[3] Compare this with Quine's "observation sentences", *Word and Object*, 1960 [11]

[4] Quine's concern relates to the commitments we might want to make in respect of the actual mind-independent existence or otherwise of the particular entities we might be referencing in propositions. This is less of an issue for everyday objects but somewhat more problematic for abstract or unobservable entities. Nevertheless we avoid getting into a discussion about 'realism', that is, whether or not the informational entities that we wish to talk about exist independently of the mind.

whether $x$ is a $P$ or $x$ is a $Q$, doesn't alter the fact, and quite probably the knowledge, that it is $x$ that he is considering. Thus we can hold that two individuals observing the same object $x$, to be in the same informational state but that they may not be holding the same content.

The knowledge required to make the distinction between $x$ and something else $y$, or to put it another way, the knowledge required to pick out a datum following the GDI definition whereby $d = (x \neq y)$ is what Evans [6] calls *discriminating knowledge* this being a subject's "capacity to distinguish the object of his judgement from all other things". In this case the concept '*knows which*' is merely the ability to notice the difference. Following Floridi [8], "the GDI is neutral with respect to the identification of data with specific relata", that is the black dot or the white sheet of paper upon which it appears, although individually distinct require each other as relational entities in order to qualify as datum, the ability we have to make this distinction would seem fundamentally critical to any understanding of information. Floridi [8] quotes Bateson who stresses the point, "what we mean by information – the elementary unit of information is a difference, which makes a difference". A difference is just a discrete state, and the making of a difference just means that the datum is potentially meaningful.

There is more to this point, at the most fundamental level we are talking about the ability to formulate a basic and primitive concept of an object (or a state of affairs). Knowledge for what it is for '$d$ is a datum' to be true necessarily, if indeed $d = (x \neq y)$, requires at least some primitive notion of what differentiates $x$ from $y$. That is by definition, we require distinguishing knowledge about $x$ since $x$ is differentiated from other objects (or states of affairs) by the fact that $x$ stands in its own right as a 'discrete state' (a datum).

According to Floridi [8] "the most important type of semantic information is *factual information*, which tells the informee something *about* something else". Implicit within the notion of factual information is the concept of identification or reference, a way of knowing which object is the one in question, summed up by Quine [9], in "no entity without identity". For only if one is able to correctly individuate or express distinguishing facts about an object from which the information derives can we assume then that that factual information does indeed say anything meaningful at all.

At this point we are merely flagging (not resolving) the issues raised by Russell's Principle as requiring a thorough reconciliation with the GDI as semantic content. The problem turns on the fact that a person's knowledge of a particular object relies upon a capacity to cite discriminating facts about the particular object from which the information derives. A denial of Russell's Principle would seemingly render a subject unable to identify a datum in the first place due to the inability to comprehend the side of the equation $(x \neq y)$ regardless of $d$. One must know there is a difference in order to ascent to as much.

## 3.

In *Phenomenology of Perception,* Merleau-Ponty [10] argues "Empiricism cannot see that we need to *know* what we are looking for, otherwise we would not be looking for

it" (italics added). Although Merleau-Ponty[5] no doubt had a much larger agenda in mind, namely to expose the problematic nature of some of philosophy's traditional dilemmas, we take his suggestion as permission to proceed in questioning prevailing *doxa*. In this section thus we explore the possibility of information and knowledge as being at least in some cases coextensive.

Coexstensivity between information and knowledge might be outlined as follows:

a) *K* (Knowledge) is a determinant of *I* (Information) if and only if *K* is functionally dependant upon *I*.
*and*
b) *I* is a determinant of *K* if and only if *I* is functionally dependant upon *K*.

In terms of the *information-first* approach to epistemology and excluding any appeals to *a priori* knowledge, ***a*** is relatively uncontroversial. I say relatively since further work is certainly required in order to thoroughly establish an information-first epistemology if indeed such an approach turns out to be correct. However more straightforwardly, ***a*** is intuitively uncontroversial compared with ***b*** because it takes for granted the popularly held assumption that an asymmetry between information and knowledge holds, whereby information naturally precedes knowledge, *as per* the continuum hypothesis.

In order to clarify ***a,*** consider the following brief example. The fact that I *know* that it is raining, for example, can be typically *determined* either by *all* the information considered relevant to such a judgment or by other information in the form of testimony provided to us by for example the meteorological office. My knowledge therefore that it is raining is *functionally dependant* upon the information germane to the fact. Such approaches appeal to our intuitive sense of information as being prior to knowledge and are typically expressed in what Quine calls *observation sentences* [11]. Information-based thoughts of this kind (observation sentences) expressed by propositions such as "it is raining", are least susceptible to variation under other informational or epistemic influences, they seem to be basic propositional attitudes.

On the other hand ***b*** is certainly controversial since it hints at a much more primordial interpretation of knowledge. Nevertheless at this point (Merleau-Ponty's concerns aside) we have identified at least one case whereby ***b*** seems to hold. The identification of any particular object or state of affairs at least requires the satisfaction of Russell's Principle. That is my making a judgment about the fact that is raining, regardless of the information I receive requires me to know what it is I am looking for. To explain further, the fact that 'it is raining' and that this is what my information based thought is focused upon requires a prior knowledge that the raw information I am immediately receiving from the environment (if indeed that is the mode by which I am receiving the information) concerns the fact that it is raining and not some other state of affairs – the judgment requires *discriminating knowledge*. As methodologically ideal as it maybe we do not operate as detached conscious beings observing and interpreting specific and discrete, epistemologically self-contained brute facts about the world.

We now introduce another case whereby knowledge appears to be in some kind of coextensive relationship with information.

---

[5] A more thorough reconciliation of Merleau-Ponty's work in regard to externalist or a "god's eye view" of information is beyond the scope of this paper, however I thank the anonymous referee for drawing this interesting connection to my attention.

Following Evans [6] an information-based thought about an object or state of affairs needs to be either a fundamental identification of that object (or state of affairs), for example 'it is raining' implies in some way that I know about rain in some essential way, or alternatively, consist in a knowledge of what it is for an identity proposition involving a fundamental identification to be true. Quite simply the necessary truth of any information-based, knowledge statement must entail the truth of the information in the first place (it couldn't be the case that 'false information' entailed any kind of knowledge), but the truth of such information surely could only be determined with a prior knowledge of what it might take for that information-based thought to be true. This is reiterated by Williamson [5] "knowing is the most general factive attitude, that which one has to a proposition". One either knows it is raining (or not), or doesn't know, or isn't sure, in the latter two cases the subject does not know at all.

A significant portion of Gareth Evan's 1982 text is committed to explaining demonstrative identification. Although we are constrained from explaining Evan's position in detail we can say that the core idea revolves around the concept of what Evans [6] calls the *information-link* between a subject and an object whereby the subject is provided with "information about states and doings of that object over a period of time". Ordinary demonstratives express informational links between objects and subjects, to the objects they identify via propositions such as 'that man', 'this chair', thus a demonstrative thought is an information-based thought.

We are not arguing that there are any identification guarantees in 'demonstrative identification' even given a clear and present information-link between object and subject. An information-link although necessary will not be sufficient to determine the truth of a demonstrative without some controlling idea or concept of the object in question. As we mentioned earlier, whether $x$ is a $P$ or $x$ is a $Q$ doesn't alter the fact that $x$ is the origin of the information, yet clearly there is a difference between a basic ability to disambiguate $x$ from other objects in the immediate vicinity and a corresponding ability to assent to the truth of $x$ being either a $P$ or a $Q$. The latter requires the knowledge of what it is for any particular object to be correctly identified as the relevant object, following Floridi, "for example where a place is, what the time is, whether lunch is ready, or that penguins are birds"[8]. Consider this in contrast with incorrectly identifying a place as some other place, or some other incorrect time, or that a penguin was some other kind of animal.

Continuing with Floridi [8], "factual information is declarative in nature [and], is satisfactorily interpretable in terms of first-order classic predicate logic". Furthermore with appeal to Dretske [2] and Barwise & Seligman [12], information carries with it the mechanism such that demonstrative identification can be satisfied "$a$'s being (of type) $F$ carries the information that $b$ is $G$". This mechanism clearly represents some kind of confirmation relation between the information-based statement and the object in question.

At this point we ought to highlight an important distinction. We can agree that information exists prior and independent to any mental processing that turns it into an information-based thought or statement (subsequently representing demonstrative identification). This is the case to the extent that 'non-thinking' organisms utilise 'information' in a behavioral kind of way, the sunflower orients itself towards the sun, single cellular organisms capable of interpreting environmental information have a survival advantage over those that are less able to interpret environmental information, likewise humans utilise various complex information channels to enable us to accomplish tasks such as riding a bike, swimming or catching a ball. This use of

information however seems to be somewhat different to our understanding and previous discussion regarding 'semantic information' in that there is clearly an absence of cognitive relations between subject and object.

On the other hand our wish is to try to establish some connections between *information* and *knowledge*, we remind ourselves of the big picture question, the speculative thought as to whether we could have an information-based epistemology. Although we don't wish to muddy the waters in respect of the ontological nature of a pure mind-independent information uninterpreted by any human mind as being somehow the same as our information based thoughts or the kind of information passed from informer to informee in the form of testimony, demonstrative identification or, 'factual information of a declarative nature' we find it difficult to discuss the one without the other. *Semantic information* is a human concept in that it involves the concepts of reference, identification and meaning, according to Colburn, "to be information requires a thinker" [13]. Such concepts of course are quite different to those of justification and belief and I see no problem in maintaining a non-doxastic, externalist position given such bounds.

Thus we have seen that there are at least two considerations whereby a coextensive relation appears to exist between information and knowledge. The first being the satisfaction of Russell's Principle and the second being an individual's ability to correctly identify objects or states of affairs as consisting in a *knowledge* of what it would take to make any proposition directly related to the objects or states of affairs in question true. Information then, it seems, carries with it what we might call a *confirmation relation*. The confirmation relation acts between the (correct) identification of objects (or states of affairs) and the agent utilising the information-link. Furthermore, confirmation relations themselves are *a posteriori* (otherwise we would have to run an argument as to how we could identify objects innately) in that they represent an empirical truth between the information-based statement (or thought) and the object to which it refers.

## 4.

The remainder of the discussion will focus on what implications *confirmation holism* might have for the philosophy of information in light of the above discussion. Confirmation holism is the premise that all theories and consequently the statements that make up those theories are underdetermined by the data or evidence supporting them [14]. The idea here is that statements or theories can never be tested in isolation. That is, we are always reliant upon other information, or knowledge in order to confirm (or disconfirm) any scientific or factual claims. Furthermore this implies that there will always be competing statements, observations or propositions that may be used to prop up a theory. Thus, in principle it is not possible to fully refute a theory based on the falsity of any individual premise encapsulated by that theory. Let us explain why this is important.

In adopting an external, information-theoretic approach to epistemology, an approach that does not demand any analysis of justification or belief in order to support its claims, we are necessarily committed to *a posteriori* or an experience-dependant account of what confirms what in the world.

The reason for this follows much of the discussion above: we are talking about mind-independent, externally generated, information-links between objects and subjects, that is, factual information of a declarative nature as representative of the confirming evidence required to make knowledge claims. Information is conveyed via perception and transmitted via language (Dummett in Floridi [8]). The factual component of an information-based statement surely must depend on the empirical confirmatory experience to which it is intimately linked. Furthermore and perhaps more importantly, the confirming methodology *itself* is also *a posteriori*. This is the case since every declarative factual informationally based proposition is itself reliant upon, firstly the knowledge that $d = (x \neq y)$ or Russell's Principle, and secondly an individuals knowledge of the meaning of any given information based proposition, following Quine [15] "the meaning of a statement is the method of empirically confirming or infirming it".

The alternative would be that confirming conditions between objects and their corresponding identifying (information-based) statements are somehow *a priori*. Such an approach ends up having to carry the burden of internalism and is thus obligated to appeals to the analyticity of confirming conditions – and the quick route back to a foundationalist analysis of justification and belief. Evans [6] echoes Quine's [15] thoughts, in that any information-based thought must be either a fundamental identification of the object in question or require knowledge of what it takes for such a proposition involving a fundamental identification to be true.

Williamson raises the issue of holism as he considers the implausibility that knowledge actually consists in "epistemically self-sufficient nuggets of information" that might exist in isolation from each other [5]. The suggestion is that if information-based statements are not somehow linked to other information-based statements in confirming knowledge claims, then they can't be confirming anything other than themselves. The *what* and *how* of any such informational connections that might exist between all the relevant information required to make knowledge claims is a story for another time; as is an answer to what mental apparatus might be required to establish such connections. Williamson [5] contributes to the suspicions in conceding to a form of holism in quoting Peirce, "I cannot make a valid probable inference without taking into account whatever knowledge I have (or, at least, whatever occurs to my mind) that bears on the question". If an information-first epistemology relies upon all the relevant information-based propositions pertinent to the particular issue as the necessary confirmation relations in respect of knowledge, then it is difficult to see how such statements could exist in isolation.

It seems that the consequence of externalism in respect to an *information-first* approach to knowledge is confirmation holism. The possession of information conveyed by perception and readied for transmission via language requires knowledge on the part of the subject, regarding the object in question, of what it takes for a proposition involving that object to be true. It further requires the satisfaction of Russell's Principle, or what Evans calls, in agreement with the Floridian definition of a datum expressed as $d = (x \neq y)$, discriminating knowledge. The discussion here does not necessarily scotch the information-based approach to epistemology, what it does do however is raise concerns over our intuitions of any presupposed asymmetry between information and knowledge.

## References

[1] Floridi, L. (2004). *Open Problems in the Philosophy of Information*. Metaphilosophy, 35(4), 554.
[2] Gettier, E. L. (1963). Is Justified True Belief Knowledge?. In *Knowledge and Belief* (Ed. A. Phillip Griffiths, 144-146). Oxford: Oxford University Press. (Originally in Analysis, Vol. 23, Blackwell pp 121-3)
[3] Dretske, F. I. (1981). *Knowledge and The Flow of Information*. Cambridge, Mass: MIT Press.
[4] Adams, F. (2004). Knowledge. in *The Blackwell Guide to The Philosophy of Computing and Information* (Ed. Floridi, Luciano. 228-236). (2004). Blackwell Publishing.
[5] Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press
[6] Evans, G. (1982). *Varieties of Reference*. London: Oxford University Press.
[7] Quine, W. V. O. (2nd Ed, Revised). (1980). *From a Logical Point of View*. Harvard: Harvard University Press.
[8] Floridi, L. (2004) Information in *The Blackwell Guide to The Philosophy of Computing and Information* (Ed. Floridi, Luciano. 40-61). (2004). Blackwell Publishing.
[9] Quine, W. V. O. (1969), Speaking of Objects, *Ontological Relativity and Other Essays*, New York: Columbia University Press.
[10] Merleau-Ponty, M., (1962), *Phenomenology of Perception*, trans, Smith, London: Routledge and Kegan Paul.
[11] Quine, W. V. O. (1960). *Word and Object*. Cambridge, Massachusetts: MIT Press.
[12] Barwise, J. and Seligman, J. (1997). *Information flow: the logic of distributed systems.* Cambridge: Cambridge University Press.
[13] Colburn, T., R. (2000). Information, Thought, and Knowledge. *Proceedings of the 4th World Multiconference on Systematics, Cybernetics, and Informatics*, 10, 467-471. Orlando, Florida:.
[14] Fodor. J. and Lepore. E. (1992). *Holism: A Shoppers Guide*. Oxford: Blackwell Publishing.
[15] Quine, W. V. O. (1953). Two Dogmas of Empiricism. in *From a Logical Point of View* (20-46). Harvard: Harvard University Press.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Phenomenal Consciousness: Sensorimotor Contingencies and the Constitution of Objects

Bastian FISCHER [a,1] and Daniel WEILLER [b]
[a] *University of Saarland/IFOMIS*
[b] *University of Osnabrück*

**Abstract.** We justify the need for better accounts of object recognition in artificial and natural intelligent agents and give a critical survey of the computational-postcomputational schism within the sciences of the mind. The enactive, dynamicist account of conscious perception is described as avoiding many problems of cognitivist functionalism, behaviourism, representationalism, emergentism, and dualism. We formalize the basic structure of the enactive, dynamicist theory of phenomenal consciousness and criticize the externalist presupposition of outside-world objects in this kind of theory. As a remedy, we suggest a sensorimotor account of objectual constitution which assigns an epistemic but not necessarily ontic priority to sense data.

**Keywords.** Phenomenal consciousness, artificial intelligence, enactivism, sensorimotor contingencies

## 1. A Desideratum for AI and Consciousness Studies

Sidney Morgenbesser is said to have criticized behaviourist Burrhus Frederic Skinner by asking him: "Let me see if I understand your thesis. You think we shouldn't anthropomorphize people?" There is a certain danger that someone with Morgenbesser's wit would ask a similar disarmingly incredulous question after being exposed to the following suggestions, which at some point do leave the natural stance of human beings. Such an adventure seems, however, necessary for properly assessing some of the difficulties that artificial intelligence and theories of the mind are currently facing. The following is to be understood under the basic premise that, from an ontological point of view, the minimally realist thesis that there are endurant, language and mind-independent individual objects in the world is not only compelling but even universally true, i.e., not only within an anthropic context. However, as far as the sciences of the mind and their accounts of conscious perception are concerned, let us briefly set aside the question whether or not this kind of weak, individual-based realism is true. However rich, parsimonious, realist, or constructivist one's ontological notion of "object" may be, it is almost impossible to cast doubt upon the claim that a few species benefitted enormously from entertaining the assumption that there are distinct, denotable objects in and around one's body.

[1] Corresponding author: Bastian Fischer, Universität des Saarlandes, Philosophisches Institut, Postfach 15 11 50, D-66041 Saarbrücken, Germany; E-mail: bafi5002@stud.uni-saarland.de.

Given this, an unbridgeable gulf appears to open between the perceptual and cognitive capacities of, *inter alia*, insect species on the one hand and many of the higher animal species on the other. It seems, therefore, worthwhile to leave the natural stance and not to take the existence of objects for granted, but to regard object recognition as the result of an evolutionary process of animals. Although our final suggestion bears some remote resemblance to Piaget's [1] constructivist notion of objectual constitution, this epistemic turn need not necessarily affect the question whether or not, say, an *ontological* realism is true.

Consider Wooldridge's description of what Hofstadter [2] called 'sphexishness' while discussing creativity. The apparently smart wasp *Sphex* usually takes a "paralyzed cricket [as food for the future wasp grubs] to the burrow, leave[s] it on the threshold, go[es] inside to see that all is well, emerge[s], and then drag[s] the cricket in" [3]. Any initial impression that she both cognizes and successfully recognizes the objects 'burrow', 'cricket' and their respective properties is, however, immediately thwarted by the following fact:

> If the cricket is moved a few inches away while the wasp is inside making her preliminary inspection, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. [3] (p.82)

While Dennett [4] appears to be claiming merely that the wasp is never *deliberately* dragging the cricket, it may also be hypothesised that she isn't even consciously dragging *a cricket* – in the sense that she would internally know that there is a distinct object in front of her. Another example of insects' cognitive acts would involve bees signalling each other by means of dancing where to find the best and nearest flowers.

It seems that these instances of reproductive and communicative behavior among insects can work well enough without their assuming that there are distinct objects such as the particular cricket or burrow or a certain meadow abounding in flowers. The wasp's hunt for the cricket, her search for a burrow and the dancing bee's discovery of brightly shining flowers and her description of how to reach them can be sufficiently explained by signal-processing mechanisms, where no sophisticated object representations are needed. To pass over the complex details of such a path description and to put it bluntly, assuming that the bee needs a rich representation of the bright, colourful meadow as the extended object *we* recognize it to be in order to tell her sisters how to find it would amount to the assumption that heat-seeking missiles need an in-built, rich representation of the warm target they are designed to destroy.

Route guidance systems work in a similar manner. They may tell a user that the spot where she, or a geometrical proxy of hers, is currently located in their map coordinates is assigned, as are some other coordinates, to the character string 'Baker Street', but such systems have no idea of Baker Street as a material object, although they do guide the user successfully from Baker Street to Piccadilly Circus. Surely enough, such systems may work backwards and let the user know all the coordinates for which it holds that, if a representation of the user's car is occupying such a coordinate, then the user is said to be in Baker Street, and, in that manner, some set or tuple of coordinates could indeed be regarded as a representation of Baker Street. The point is, however, that this is simply what the system has explicitly been told in one way or another. The system in a way *knows* Baker Street, but never *actively got to know* it.

We believe, however, that *our* (and quite plausibly some other animals') cognitive processes involve more than that. Imagine person *A* is lying sick in the bed and would like her partner to get her an antipyretic or a glass of water from the kitchen. In this scenario, it would be significantly uncomfortable for *A* to guide her partner by an exact description of respective location coordinates and/or move instructions to a place called kitchen, a thing called glass, tap, etc., and this even if the mechanisms of grabbing the glass and turning the water on were already dispositionally implemented in *A*'s partner without the need of any sophisticated object representations. Luckily, *A*'s partner is usually in a position to recognize even different types of objects and pick out similar instances of 'glass', 'kitchen', 'antipyretic', 'tap', etc.

Human intelligence involves, at least *prima facie*, the assumption that there are objects in the material world as well as the ability to regard certain chunks of this material world as *distinct* objects. We conclude that every viable account of any kind of intelligence, consciousness, or, generally, mind which is to be on a par with human intelligence, human consciousness, or human mind should be in a position to explain just that assumption and just that ability of object recognition. Ideally, it should also be able to explain how an agent can make that assumption and can recognize objects *as* objects *on its own* – independently of a designer or programmer. For instance, the designer or programmer of an *artificial* intelligent agent should, sure enough, implement the ability of automatic object constitution but not explicitly – and from her *external* programmer's viewpoint – let the artificial agent in on what the exact objects to be recognized are.

## 2. The Sciences of the Mind

This ideal desideratum of both AI and accounts of animal consciousness being stated, what follows is a critical survey of the situation that the sciences of the mind currently find themselves in. There is arguably some schism between computational and post-computational aspects of the mind reflecting a similar schism of general approach in the sciences of the mind. Apart from other issues, the question arises whether or not the estranged halves in both types of schism may ever be soldered together. We will not claim that such soldering could be undertaken without the risk of researchers encountering serious difficulties – and not least also with respect to a coherent philosophy of science – but we do claim that, in principle, such a union could be realized and that it promises progress. Our final outline of the structure of dynamical object-constitutive processes may be regarded as an attempt at making post-computational aspects of the mind computationally accessible.

On the one hand, there is the influential and still somewhat main-stream *cognitivist* paradigm according to which functionalism or sophisticated mind-brain identity theories are, in a way, to be regarded as a satisfactory, full-fledged philosophy of mind (cf. [5]). The brain speaks a computational language of its own; and the conscious experience of, say, seeing a red tomato can be equated with a certain functional system state of the brain – functional in so far as it fulfils the causal roles of the conscious experience, i.e., it is a state that is caused by the presence of a red tomato in the receptive field and that causes our mouths to water and/or another mental state such as the wish to eat the tomato (or such as disgust if you don't like tomatoes) [6]. In this human case, the information processing system detecting the tomato is seen as physically implemented in a human body with a special focus on the brain. Furthermore, the information processing system is considered as highly decomposable

into basic units some of which represent the outside-world entities that are being perceived. These units, called representations, are as parts of the information processing system (besides the actual processes transforming information) also physically implemented in the human brain. Ergo, representations are physical entities carrying information [7].

Along these lines, it has been hypothesised that, for instance, heights of rectangles could be physically represented in the neural information system by certain firing rates of neurons, where these neurons fire at a rate proportional to the height of the rectangle perceived [7]. In this example another basic hypothesis of representationalism is expressed: the theory of isomorphism. It is claimed that there exists an isomorphic mapping between outside-world entities and internal neural representations, i.e., a bijection between the two sets of objects such that relations among the external objects are reflected by corresponding relations among those internal representations that the external objects are bijectively assigned to. We should note that this gross explication of isomorphism has undergone some refinement: There are, firstly, non-topographical versions of isomorphism that do not require that any local relations "outside" be mirrored by local relations "inside" [8]. Secondly, gross isomorphism is often replaced by what Noë & Thompson call the *matching content doctrine.* Here, the isomorphism is confined, in a somewhat obscure way, to the representational *contents* of the neural system and the *contents* of conscious experience [8].

We should also note that many versions of today's mind-body dualism are fairly compatible with the representationalist tenet of functionalism, which can assume quite a reductionist, physicalist form [6]: David Chalmers [9], on his search for a neural correlate of consciousness, posed several preliminary questions, two of which are: "How can we find the neural correlate(s) of consciousness?" and "Is consciousness reducible to its neural correlate(s)?" Here, he does not ask "Are there neural correlates of consciousness?" but, through the use of the definite article and the success verb "find", presupposes that NCCs do exist. As the second question of reducibility of consciousness to its neural correlates is not idle and self-answering, it is plain that the theory of representationalism is by no means incompatible with those antireductionist, antimaterialist theories of mind that Chalmers (e.g., [10]) so eagerly defends. On his search for NCCs, he is looking for a neural system $N$ such that there exists a bijection between neural states $n$ of $N$ and conscious states $b$ (of a person), where the occurrence of a given state $n$ of $N$ is a minimally sufficient condition for the occurrence of the corresponding conscious state $b$ [9].

The bijective mapping, we dare say, could be explicated not only reductively by primitive identity relations but also by psychophysical laws typical of substance dualism or epiphenomenalism – the view that consciousness plays the role of a mere varnish on the physical realm, caused by the latter but irreducible to it and in itself causally inert. Even Chalmers's neutral monism is not out of the game: A neural state $n$ of $N$ may simply be assigned to a conscious state $b$ iff both $n$ and $b$ (or their essential properties) are constituted by *protophenomenal* properties of one and the same fundamental physical entity (for instance, by the protophenomenal properties of the brain). Here, the *protophenomenal* properties are those which Chalmers regards as constitutive of the neutral third type of entity on which both the mental and the physical are grounded.

We give two final examples for a hardcore representationalist theory. It has been suggested that certain neurons in a monkey's inferotemporal cortex represent butterflies (because they fire when the monkey is consciously experiencing the image of butterflies; [11]); and according to Barlow [12], some specific neurons could, in principle, be able to represent none other than your grandmother.

This form of representationalism has been criticized not only by proponents of emergent property theories according to which global states of the brain can carry meanings beyond the sums of the meanings of the components of the brain and according to which cognitivist computer models of the mind will, bare and as such, never penetrate the so-called sub-personal level of the goings-on of mind and reach the personal level of experience (cf. [5]). Representationalism has also come under attack by proponents of the so-called *enactive* paradigm of cognitive science, especially by O'Regan & Noë's [13] sensorimotor account of vision and visual consciousness as well as by Noë & Thompson's [8] negative answer to the question "Are there neural correlates of consciousness?". The arguments for this criticism are both empirical and philosophical.

Some neuropsychological findings are hardly compatible with the existence of static, single, internal representations of objects. Consider the observation that characters as of an unknown writing system can hardly be learned, recognized, and distinguished by a subject if their perceptual apparatus is still and fixed on one central point when being presented with the characters [14], which implies that a dynamic plurality of different signal registrations is necessary for the perception of entities. Consider the problem of the blind spot recently revived (e.g., [15-17]). Certain parts of the outside world that lie within the visual field cannot possibly be represented at all times in the brain if the light waves reflected by them hit the blind spot on the retina, yet the visual field appears to have no hole or distortion. O'Regan [18] argues that there is no need for a mechanism filling in the blind spot. Although vision researchers (e.g., [19]) have uncovered that certain retinal or cortical mechanisms may bear a direct relation to a subject's experiencing how a part of the visual field obstructed by the blind spot appears continuous, there is reason to doubt that these mechanisms really serve to fill in an internal, picture-like copy of the outside world. O'Regan & Noë [13] caricature this view by referring to the spatiotemporal integration in the low-level visual system, which explains why very closely spaced dots look like lines, and stress that no one would claim here that this integration mechanism actually fills in the gaps between the dots. Most importantly, consider what happens in a person's visual cortex when she looks directly at a straight line and then looks above that line, so that she sees the line at a lower point of her visual field than before. In the first stance the cortical activation pattern will be fat and lentil-shaped whereas it will be meagre and banana-shaped when the subject is looking above the line [13]. Were the line to be represented in the brain by a static pattern, then some mechanism would have to be invoked which would transform the banana into the lentil or vice versa, depending on which pattern should be the "true" representation of the line to be processed by the rest of the brain. This hypothesis is, however, as implausible as a certain mechanism actually filling in the blind spot.

It might be objected that none of these empirical findings provides a fatal argument against representationalism because the notion of representation that cognitivism is based upon may easily be refined so as to accommodate the phenomena just mentioned. We fear, however, that, once such refinement is elaborated, the result will at least go against the essence of cognitivism. To accommodate the first phenomenon, one will have to introduce dynamically shifting representations that nonetheless stand for a single entity. In the last case, given that there are obviously plenty of interstages between lentil-shape and banana-shape, one will even have to accept a vast, perhaps infinite multitude of possible representations for a single line. Such an endeavour seems to miss the point of cognitivism, which aims at fine-grained, semantic decoding of brain structures.

To draw an analogy: In the theory of neural correlates of consciousness, the view that not less than the whole brain is any such NCC is commonly rejected on the basis of the same fundamental cognitivist goal (cf. [9]); although the brain as a whole does trivially fulfil the condition that it is a neural system *N*, where the occurrence of a given state of *N* is arguably a sufficient condition for the occurrence of a corresponding conscious state – albeit not a minimally sufficient one. It may be true that the multitude of neural processes in the whole brain appears much greater than the mulititude of the goings-on in the visual cortex during the dynamical visual perception of the line. The latter multitude seems, however, still much too large and unmanageable when decoding the semantic content of the brain. Therefore, one might wonder whether cognitive scientists in favour of such a vast set of possible representations of a line might not as well adopt the view that not less than the whole brain is any NCC. The objection that each of the single, distinct states in the visual cortex is at least a *minimally* sufficient condition for the occurrence of the visual experience of the line is somewhat corrupted by the observation of the first empirical example, namely of the fact that single, momentary states will not suffice for the description of a conscious experience but that, instead, a plurality of such states must be considered. Once, however, such a plurality of states enters the theory of consciousness, we would have to deal with dynamical representations of the entities perceived. Noë & Thompson [8] regard the existence of such representations as an interesting possibility. They remark, however, that dynamicist neuroscientists (e.g., [5]) usually "shun talk of 'internal representations'" and hold that cognitive processes "span the nervous system, the body, and the environment" – a thesis that Thompson & Varela [20] extend to *conscious* processes.

In our view, the introduction of a dynamical kind of representation will fail exactly because of this wide local extension of cognitive as well as conscious processes: A plausible and practical way of dynamically explaining a certain conscious experience of an object will allow a designator of the perceived object itself to figure in areas of the *explanans* where, according to representationalist cognitivism, only designators of (neural or other) representations of the object would seem suitable; and, according to a standard view, objects do not represent themselves. On the enactivist account, an outside-world object, of which the subject can have explicit knowledge, may have the same explanatory status as some neural entity of which the subject can have implicit knowledge (see below).

But let us turn to a philosophical argument against representationalism and isomorphism: Some claim that the contents of consciousness and the measurable contents of the receptive field of a subject are incommensurable [8]: The content of a conscious experience is structurally coherent, the content of the receptive field is not. You cannot imagine the experience of a complete butterfly without the experience of a background where the butterfly is situated. The signals of the receptive field, however, are decomposable into elements of background and of butterfly. Emergentists try to resolve this so-called binding problem by saying that, for instance, the representations of different properties (of background, butterfly, shape, colour) that are implemented in different neural pathways are tied together in the brain as a global system by synchronised firing rates between the brain parts involved; and it has been found indeed that such synchronised firing rates are correlated with the perception of gestalt or grouping criteria [21]. However, given that the conscious experience is, by phenomenological reduction, if you wish, structurally coherent at all times, the binding problem is simply malformed: The content of the conscious experience of the butterfly is simply not composed of static representational inscriptions of features and properties of the content of the receptive field and can thus not be equated with a binding together of various such inscriptions in the brain. Rather, the structurally coherent content of the

visual experience is situated in the subject's egocentric space – but it is not clear how a neural representation could possess a content that is situated in egocentric space. Noë and Thompson:

> Although neural systems causally enable the animal as a situated agent to orient itself in its egocentric space, they themselves do not inhabit this space, nor do they have any access to it as such" ([8], p.15).

Human beings and other animals "experience the world as laid out before them, but the neurons do not" ([8], p.16).

## 3. Sensorimotor Contingencies and Objects

Now, is there any way out of these problems? As we have seen, emergentism is no perfect solution, and even to become a dualist does not save one from the difficulties of representationalism. A potential solution for the philosophy of mind – apart from eliminative materialism – is O'Regan & Noë's [13] sensorimotor account of consciousness: They offer a fairly materialist account of *sensorimotor contingencies* which is significantly more than behaviourism and functionalism. Their theory is enriched by the concept of practical knowledge about the changes of sensual stimuli that dynamic changes of the perceptual apparatus relative to the perceived object bring about. These changes follow the laws of nature. Behaviourism does not involve such knowledge about one's own conscious states or processes. Functionalism, in principle, allows for such knowledge, but, as has also been said, functionalism usually assumes single and static representations within those system states that fulfil certain causal roles in a conscious system and that give rise to, or are equated with, conscious states. When utilizing *sensorimotor contingencies*, however, there must occur at least two different physical states of the conscious system in order to account for a conscious experience.

We will briefly outline the first steps towards a formalization of O'Regan & Noë's theory. This might help when implementing their model into artificial cognitive agents.

According to O'Regan & Noë, perception is a way of exploring the world that is mediated by practical knowledge about sensorimotor contingencies. These are the laws that govern the changes of sensual stimuli that dynamic changes of the perceptual apparatus relative to the perceived object bring about. The concept of *practical* knowledge can, as a first attempt at formalization – and given an explication of the concept of knowledge – be explicated as follows:

> (PK) Subject *S knows practically* that $v \rightarrow z$ *(material conditional)* iff *S* knows that $v \rightarrow z$, where *v* describes the execution of a dynamic process of the perceptual apparatus relative to the perceived object, and *z* describes a new state (or process) nomically resulting from the execution of that process described in *v*.

*v* and *z* are propositional or sentential variables. Here, for the perception of a chair, by *v*, such simple propositions as "I move closer to the chair" may be meant, and, by *z*, a proposition like "I am right next to the chair (as opposed to ten metres away from it as before moving)" or "the chair touches my leg" or "some of my haptic nerves from my

leg fire". The last example gives an idea of the equal explanatory status of internal and external objects in an enactivist explanans of the explanandum 'conscious experience'.

The experience of a specific perceptual object is marked off as a class from all other experience classes by stipulating: The experience of a specific perceptual object is the set of all potential, dynamic, and law-like changes of the perceptual apparatus relative to that perceived object (where, say, a *visual* experience is differentiated from *hearing* with the help of certain contingencies such as: if the subject covers her ears, the sense input will not change much, if she covers her eyes, it will change significantly). The experience of a certain object is then instantiated in subject *S* iff

(i) at least one dynamic process of the perceptual apparatus relative to the object described in a *v* is actually executed (i.e., there exists a *v* that is true), where there is at least one *z* such that *S* knows that *v* and practically knows that $v \rightarrow z$,

(ii) the practical knowledge is integrated in *S*'s planning of behaviour, rational judgement, or use of language (the so-called awareness-condition).

(i) may, of course, be reduced to "there is at least one *v* and at least one *z* such that *S* knows that *v* and practically knows that $v \rightarrow z$" because knowledge that *p* implies (the truth of) *p* and because the inner structure of what is described by *v* can be derived from the fact that *S practically* knows that $v \rightarrow z$, i.e., from (PK). So, if you are perceiving a chair, this may mean (i) that you move closer to the chair, know that you move closer to it, and know that, if you move closer to it, the chair will, say, touch your leg, or, say, (due to the larger retinal image of the chair) many more retinal cones will be stimulated by light reflected from the brightly red backrest than were before moving, both of which is *practical* knowledge, and (ii) that, for example, you plan to sit down on the chair, apply a simple test of its stability to it, or utter: "Oh, nice chair over there!" – all of which Chalmers would, in his famous dichotomy between "hard" and "easy" problems of consciousness [10], classify as "easy" ones.

It might be objected that processes such as those described by *v* may only reasonably be analysed as ordered static snapshots such as "I am 10m away from the chair", "I am 9m away from the chair", ..., "I am right next to the chair" so that the first *z* suggested above would be included in *v* and so might trivially follow from it; however, this problem can be avoided by requiring that *z* involve relations not yet stated in the description of the *v*-process. It might also be objected that practical knowledge involving resulting states/processes internal to the perceptual apparatus such as "my haptic nerves from my leg fire" or knowledge about retinal processes is rather implicit than explicit. It can externally be assigned to the subject in a certain sense, but she herself has no idea of it. This is, however, merely a contingent fact; a person permanently equipped with the best and handiest neuroimaging techniques could, in principle, turn all this implicit knowledge into explicit knowledge such that knowledge about internal processes could become as explicitly known as the fact that your hand will finally touch the chair when you keep moving it in its direction.

This seems fine, but what about the objects? The sensorimotor account involves no need for inner objects in the sense of static representations. Nonetheless, O'Regan and Noë clearly speak about the external objects that are being perceived.

In the framework of this account, the existence of objects in the environment of the subject is externally presupposed. Indeed, as Wright ([13], (commentary) p.1010) puts it, O'Regan & Noë view objects somewhat "magically" as "already given units in the external real." In its current state, therefore, the theory cannot, if implemented, enable a cognitive agent to constitute material objects automatically (or explain how we as humans first came to believe that there are distinct objects in and around us). The ideal

requirement for modelling intelligence which would be on a par with human intelligence is not fulfilled. From a functionalist, representationalist point of view, this problem of object constitution or recognition was, in theory, not that serious because for modelling intelligence and object recognition it was simply required to provide an artificial agent with little sisters and brothers of outside objects, i.e., with the representations of these objects. *We* simply have them in *our* heads and so it should be no problem to get them into other heads or artificial systems. But now that we see that not even we have isomorphic, static, simple representations of the outside world in our heads, the problem of object recognition appears in a whole new light. So, a proposal for an extension of O'Regan & Noë's theory that takes care of this further desired step in the development of understanding and modelling human consciousness (i.e., an account of how objects epistemically become what they are, how they are recognized) is needed.

Our proposal for a solution would be the use of dynamic constitutive processes for objects. They will involve a finite series of potential object variables. Such a variable is assigned a value whenever a certain input of similar, repeated sense data occurs. Thus, the enactive objectual constitution of a non-moving staircase may, roughly and two-dimensionally speaking, involve the repeated input, into a randomly moving sensing device, of signals emitted from marked dots (*) in the vertices of the steps in the following manner: *(0|0), *(0|1), *(1|1), *(1|2), *(2|2), *(2|3), *(3|3), *(3|4), etc., where (x|y) are location coordinates relative to a fixed point. Notice, however, that the sheer similarity or repetition of such sensing and respective distance measurement should make the agent assume an object. It is not explicitly told that the given pattern *conditional on similar patterns of dynamic movements* is a staircase (which it isn't, it does not even represent one) but it may give it a random name in its own language – along with assigning to the newly found object a location as determined by the distance measurements[2] – before, perhaps, learning that humans, for example, talk about "a staircase" when discovering similar sensorimotor contingencies.

This pattern simply figures as the value of some variable in a memory of "objects". Once the agent recognizes that, conditional on certain movements, a sequence of sense data it receives has already come in several times before, such a sequence becomes the value of such an object variable. Rather than externally ascribing to this variable as a meaning the real-world object from which the signals are emitted, its semantic content will be confined to the sensorimotor contingencies the sentient subject just discovered plus a location – that is, to the similar sequences of sense data *conditional on similar sequences of dynamic movements* relative to a fixed point and a location derived from distance measurement (which, luckily, does not rely on the objectual character of the entity from which the distance is measured). If the artificial machine is provided with the right types of sense and with appropriate innate behavioural goals and desires, the objects it will constitute could be quite similar to the ones that human beings and many other animals assume to exist. We do admit that this enactive account of object constitution assigns an epistemic priority to sense data[3], which, however, need not necessarily reflect an ontological priority, and that such an account might presuppose space and perhaps time as primitively real physical dimensions that are as such

[2] These may be achieved with the help of lasers, ultrasound, or in the primitive human manner by simply running into the object and counting the steps or arm lengths or seconds before contact.

[3] We understand the tem "sense datum" in quite a loose sense, not, for example, the strict subjectivist, mind-dependent one common in the philosophy of perception. In our view, an automatic colour detector may have a "sense datum" of red once it is in a position to detect radiation of a certain wavelength in the electromagnetic spectrum.

irreducible to relations among objects, albeit ontologically dependent upon them. But if this is granted, the epistemic birth of objects seems no irreducibly anthropic mystery.

## Acknowledgements

We thank Darren Abramson, Johanna Seibt, Marcin Milkowski, and Timothy Mann for comments and discussion and Peter König for inspiration.

## References

[1] Piaget, J. (1970) *Genetic Epistemology.* Columbia University Press.
[2] Hofstadter, D.R.. (1982). 'Can creativity be mechanized?' *Scientific American* 247, September: 18-34.
[3] Wooldridge, D. (1963). *The Machinary of the Brain.* New York: McGraw Hill: 82.
[4] Dennett, D.C. (1995). *Elbow Room: The varieties of free will worth wanting.* Cambridge, MA: MIT Press.
[5] Varela, F.J., Thompson, E. & Rosch, E. (1991). *The Embodied Mind.* Cambridge, MA: MIT Press.
[6] Block, N. (1999). 'What is functionalism?' in Block, N. (ed.) *Readings in the Philosophy of Psychology*. London: Methuen. 171-184.
[7] Palmer, S.E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.
[8] Noë, A. & Thompson, E. (2004). 'Are there neural correlates of consciousness?' *Journal of Consciousness Studies* 11, No. 1: 3–28.
[9] Chalmers, D. (2000). 'What is a neural correlate of consciousness?', in Metzinger, T. (ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Cambridge, MA: The MIT Press/A Bradford Book.
[10] Chalmers, D. (2003). 'Consciousness and its place in nature' in: S. Stich and F. Warfield (eds.) *Blackwell Guide to the Philosophy of Mind* (Blackwell, 2003) and in D. Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings* (Oxford, 2002).
[11] Sheinberg D.L. & Logothetis, N.K. (1997). 'The role of temporal cortical areas in perceptual organization' *Proceedings of the National Academy of Sciences USA* 94: 3408–13.
[12] Barlow, H. (1972). 'Single units and sensation: A neuron doctrine for perceptual psychology' *Perception* 1: 371-394.
[13] O'Regan, J.K. & Noë, A. (2001). 'A sensorimotor account of vision and visual consciousness' *Behavioral and Brain Sciences*, 25 (4): 883–975.
[14] Nazir, T.A. & O'Regan, J.K. (1990). 'Some results on translation invariance in the human visual system' *Spatial Vision* 5(2): 81–100.
[15] Ramachandran, V.S. (1992) 'Filling in gaps in perception: I.' *Current Directions in Psychological Science* 1(6): 199–205.
[16] Ramachandran, V.S. (1995) 'Filling in gaps in logic: Reply to Durgin et al.' *Perception* 24(7): 841–45.
[17] Durgin, F.H., Tripathy, S.P. & Levi, D.M. (1995). 'On the filling in of the visual blind spot: Some rules of thumb' *Perception* 24(7): 827–40.
[18] O'Regan, J.K. (1992). 'Solving the "real" mysteries of visual perception: The world as an outside memory' *Canadian Journal of Psychology* 46(3): 461–88.
[19] Paradiso, M.A. & Nakayama, K. (1991). 'Brightness perception and filling-in' *Vision Research* 31(7–8): 1221–36.
[20] Thompson, E. & Varela, F.J. (2001). 'Radical embodiment: Neural dynamics and conscious experience' *Trends in Cognitive Sciences* 5: 418-25.
[21] Engel, A.K., König, P., Kreiter, A.K., Schillen, T.B. & Singer, W. (1992). 'Temporal coding in the visual cortex: New vistas on integration in the nervous system' *Trends in Neurosciences* 15: 218–26.

# Part IV

# Computing in Society: Designing, Learning, and Searching

This page intentionally left blank

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Towards an Intelligent Tutoring System for Propositional Proof Construction

Marvin CROY [a,1], Tiffany BARNES [b] and John STAMPER [b]
[a] *Department of Philosophy, The University of North Carolina at Charlotte*
[b] *Department of Computer Science*

**Abstract.** This article reports on recent efforts to develop an intelligent tutoring system for proof construction in propositional logic. The report centers on data derived from an undergraduate, general education course in Deductive Logic taught at the University of North Carolina at Charlotte. Within this curriculum, students use instructional java applets to practice state-transition problem solving, truth functional analysis, proof construction, and other aspects of propositional logic. Two project goals are addressed here: 1) identifying at-risk students at an early stage in the semester, and 2) generating a visual representation of student proof efforts as a step toward understanding those efforts. Also discussed is the prospect for developing a Markov Decision Process approach to providing students with individualized help.

**Keywords.** Proof construction, intelligent tutoring system, Markov decision processes, visualization

## Introduction

Intelligent tutoring systems generally involve three component models: one of the student, one of the teacher/pedagogical technique, and one of the subject matter [1]. Each model presents its own challenge in respect to construction and effective use. One issue in respect to student models concerns their content, i.e., what student characteristics should be included, what aspects of performance should be represented, etc. One challenge in respect to pedagogical technique concerns the form and timing of help [2]. Perhaps the most challenging task concerns the explication of expert knowledge concerning the subject matter and how this knowledge relates to pedagogical technique. The number of hours required to achieve this task are normally prohibitive and certainly constitute one of the most costly components of an intelligent tutoring system [3].

[1] Corresponding Author: Marvin Croy, Department of Philosophy, University of North Carolina at Charlotte, Charlotte, NC, 28223, USA.

The activities reported here address a number of these challenges in respect to teaching deductive proof construction in propositional logic. In respect to student models, our aim is to discover variables that will identify the most at-risk students at an early stage in the course. Once we can reliably predict which students might need special help, the challenge is to design activities that will provide this individualized assistance. One objective is to supply individualized help as students use instructional programs that facilitate the learning of proof construction. The general nature of this assistance is guided by visual representations of student problem solving efforts, and by a novel approach to leveraging past student data to provide intelligent assistance. This approach is based on using Markov Decision Processes (MDPs) to represent student paths connecting premises and conclusion [4].

## 1. The Course Curriculum

Our Deductive Logic serves general education students during their first two years of undergraduate study. This course begins with state transition problem solving. Students practice a number of classic problems such as the Towers of Hanoi, Water Jugs, and Jealous Husbands [5]. In this context a problem is defined in terms of a starting state, goal state, and transition rules. This readily maps on to deductive proof problems via a premise set (starting state), a conclusion (goal state), and a set of valid forms of inference/replacement (transition) rules [6]. In particular, working backwards from goal to starting state translates into working from conclusion to premises, and this provides an alternative strategy for discovering proofs [7,8]. The mastery of proof construction and valid inference-making is central to the course. The skills developed in proof construction are made use of in subsequent course topics, such as SQL-type database searching, decision making in the context of structured documents (e.g., executing Internal Revenue Service tax form instructions), and argument analysis in natural language. This point is crucial since failure to master deductive inference and proof construction becomes evident when students reconstruct natural language passages via deductive patterns of inference.

Students face examples similar to the example shown in Table 1. Students are expected to recognize premise/conclusion relationships, reformulate statements, explicate implicit premises, identify assumed synonyms, delete unnecessary information, and fit the argument to a valid form where possible (which sometimes requires the application of rules of replacement). In particular, when students discover proofs by working backwards from conclusion to premises, they postulate sub-goal expressions that, if ultimately justified, lead directly to the proof's conclusion. Mastering this process facilitates the ability to identify implicit premises in natural language arguments. In the example given in Table 1, notice that the second premise ("If x can be attained only through the methods of the natural sciences, then x is based on generally observable facts") is not given in the original argument and must be explicated by the student. Familiarity with the rule of Hypothetical Syllogism aids in making this explication. While the argument given for reconstruction may seem trivial, this is exactly the kind of problem that weaker students fail to complete.

**Table 1.** Sample argument and components of its reconstruction.

| Argument and Reconstruction | Basis for Reconstruction |
|---|---|
| "Finally there is the triumphant idea of positivism, that valid knowledge can be attained only through the methods of the natural sciences and hence that no knowledge is genuine unless it is based on generally observable facts" [9]. | **1) Distinguish premise(s)/conclusion**<br>(conclusion indicator = 'hence' )<br>**2) Fill in information**<br>(filled in referent for 'it')<br>(supplied implicit premise)<br>**3) Delete information**<br>(Deleted 'Finally there is the triumphant idea of positivism, that' on basis of being irrelevant to the inference pattern)<br>**4) Reform statements into conditionals**<br>(universal subjects / necessary conditions)<br>**5) Assumed Synonyms**<br>('valid knowledge' = 'genuine knowledge') |
| Given premise: If A then B<br>Implicit premise: If B then C<br>Conclusion: If A then C | **A** = 'x is valid knowledge'<br>**B** = 'x can be attained only through the methods of the natural sciences'<br>**C** = 'x is based on generally observable facts' |

Throughout the course students make use of instructional Java applets[2]. These programs not only provide opportunities for students to develop and hone relevant skills, but they also support the timely collection of data on student performance. This is particularly important during the early weeks of the semester. During this period, students use applets to complete state transition problems, carry out truth functional evaluation, and practice the application of deductive proof rules. Two applets are central to this curriculum. "Justified Thought" (JT) provides practice with the inference/replacement rule set. JT uses mal-rules and complex instantiation to build a deep understanding of rule applications. The program provides three levels of increasing difficulty for both inference (implicational) and replacement (equivalence) rules. Students must judge whether particular expressions do or do not conform to various rule patterns. In level one, some rule pattern always fits the given expression, which JT constructs by instantiating rule forms with simple statement letters. In higher levels of JT, some expressions fit none of the rule patterns shown. In these cases, the expression fits a mal-rule, a deviant of the actual rule. Moreover, higher levels of JT use complex expressions to instantiate rule variables.

Practice with JT occurs in parallel with work in "Deep Thought" (DT), a Java applet that provides a graphic environment for building proofs. DT supports both working forwards and backwards and, along with JT, maintains detailed records of student efforts. Developing prowess in constructing proofs provides the main challenge of the course. When investigating the difficulties students have in learning proof construction, attention is naturally drawn to two main skills: rule application and

[2] Students access these applets via a course management system such as BlackBoard that supports applet use by means of a number of additional resources. Direct, unsupported access to the applets is available via http://itsxserve.uncc.edu/philosophy/tallc/applets/applets.html.

strategic planning. Logic textbooks almost invariably isolate these activities by presenting rule application exercises prior to full proof problems. Nevertheless, it should be understood that the selection and ordering of rule applications can be shaped by strategic considerations. One question of interest is how rule application and strategy selection interact, whether positively or negatively. Obviously, there is a positive interaction when strategy selection (e.g., breaking expressions down into components) suggests the application of various rules (e.g. Simplification). However, when strategic thinking proposes a highly useful intermediate expression that could lead to the conclusion, similarities between this proposed sub-goal expression and some premise (or previously derived expression) may produce attempted misapplications of rules. For instance, when 'not A' constitutes a goal expression, students are temped to misapply the rule of Simplification to the premise expression 'not (A and B)'. Here, the main connective of the premise expression is actually a negation and not a conjunction as required for application of Simplification.

Another aspect of our efforts concerns the question of what contributes to proof problem difficulty and how these conditions relate to the connection between proof construction and argument reconstruction. As suggested in the sample argument in Table 1, argument reconstruction often involves the removal of irrelevant information. The analog within proof construction is superfluous premises. It is unfortunate that logic textbooks do not routinely present proof problems with unnecessary premises. Particularly when the superfluous premises contain components similar to the components of other premises or the conclusion, problem difficulty increases, and the required distinction between the relevant and the irrelevant provides part of the bridge to argument reconstruction. Our visualization method helps identify other problem areas in proof construction that we can later analyze for their impact on argument reconstruction.

## 2. The Search for Early Predictors

The first data set analyzed includes records from 85 students enrolled in three sections of Deductive Logic during 2007. Data from these students include: 1) scores on a pre- and post-test (number correct out of 25 multiple-choice items, primarily focused on validity judgments and deductive inferential tasks), 2) performance measures from a state-transition problem (Jealous Husbands: number of incorrect moves), 3) success rate for level 3 performance on truth functional evaluation of complicated logical expressions, 4) success rate for levels 1, 2, and 3 of the JT rule practice applet, 5) two exam scores (mid-term and final), and 6) course grade (expressed as percent of total possible points on exams plus additional assignments).

In order to correctly identify at-risk students early in the semester, we searched for items from our data set that could provide a significant correlation with exam scores and final course grade. Pearson correlations were run using SPSS statistical software to analyze the correlation between the exam scores, final course grade, and the following predictors: pre-test score, Jealous Husbands Puzzle success (measured by number of solution steps and incorrect moves), Truth Functional Analysis success rate, and Justified Thought success rates for levels 1, 2, and 3. The resulting correlations can be seen in Table 2.

**Table 2.** Correlations of potential predictors with course performance.

| *Course Performance* | *Pre-test* | *Jealous Husbands Puzzle* | *Truth Functional Analysis* | *JT Rule Exercise Level 1* | *JT Rule Exercise Level 2* | *JT Rule Exercise Level 3* |
|---|---|---|---|---|---|---|
| Exam 1 | .359 (*) | .072 | .238 (*) | .440 (**) | .484 (**) | .617 (**) |
| Exam 2 | .129 | -.413 (**) | .198 | .525 (**) | .447 (**) | .302 (*) |
| Exam Total | .271 | -.196 | .249 (*) | .566 (**) | .533 (**) | .541 (**) |
| Course Grade | .242 | -.305 | .173 | .573 (**) | .559 (**) | .491 (**) |

* significant at .05 level
** significant at .01 level

From the results, there are two interesting points that immediately stand out. First, the pre-test is not the most reliable predictor for success in the class. Although pre-test performance had some correlation with exam 1 (significant at p=.05, but not at p=.01), it did not correlate well with Exam 2 or final course grade. This result was surprising, since we originally felt the pre-test would predict class performance. However, the lower correlation between the pre-test and the final course grade is a healthy indicator that learning can occur and student performance does not depend entirely on prior preparation.

The other main finding from these statistics is how well the JT success rates correlate with exam scores and course grade. Based on this, we created metrics using the JT success rates as our early predictors. The goals of our metrics were to identify as many at-risk students as possible while minimizing the number of students incorrectly assigned to this category. To further this investigation, we selected a past class (Spring 2007, with 30 students) to serve as a starting case (training set), and the most recent class (Summer 2007, with 20 students) as a test case (trial set). In the training set, low course performance was defined in terms of a course grade of 65% or lower. We determined cutoff values for each of the three levels on JT (JT1, JT2, and JT3), and examined how well they identified low-performing students.

The extent to which various success rates on JT correctly identified students with low performance is shown in Table 3. This table shows the number of students who were correctly classified as low-performing (course grade of less than sixty-five percent) according to five different metrics based on JT success rates: less than 90% on JT1, less than 70% on JT2, less than 75% on JT3, a combination of each of these three, and a combination of only JT2 with JT3 performance. Table 4 also shows the number of students misclassified by each of these metrics. Misclassification occurs as false negatives (students who scored above the metric success rate but who were actually low-performing) and as false positives (students who scored below the metric success rate but who were not low performing). The results show that the best results occur for a combination of JT2 with JT3. This metric correctly classifies 25 students, generates no false negatives, and identifies only 5 false positives. False positives, students inaccurately identified as being at-risk, are not particularly troublesome, since these students would likely be given more assistance than actually required. False negatives, at-risk students treated as not being so, are of more concern, and their low numbers here are encouraging.

**Table 3.** Number of low-performing students identified by JT performance levels (Training Set).

| | *JT1 Success Rate < 90%* | *JT2 Success Rate < 70%* | *JT3 Success Rate < 75%* | *JT1, 2, & 3* | *JT2 & JT3* |
|---|---|---|---|---|---|
| Correctly Categorized | 22 | 27 | 22 | 23 | 25 |
| False Negatives | 5 | 2 | 3 | 0 | 0 |
| False Positives | 3 | 1 | 5 | 7 | 5 |

The results of testing these metrics against data from the most recent logic course are shown in Table 4. Once again, many more correct than incorrect classifications are made using any of the previously defined metrics, and most misclassifications are false positives. However, the metric that combines performance on JT2 with JT3 was slightly outdone by the metric that focuses only on JT2. In sum, the metrics defined in the training set do a good job of classifying students in ways that, from a pedagogical standpoint, reliably identify at-risk students. The best of the JT metrics serve to predict students who are candidates for special attention or assistance. Answering questions about the nature of that assistance now becomes even more crucial.

Clearly, the first step in addressing at-risk students is to ensure that these students complete practice with JT, and particularly JT2, with at least 70% correct. However, the next step in assisting students is not as clear. In the next section, we discuss one alternative for providing proof hints within DT that will specifically benefit at-risk students.

## 3. Visualizing the Nature and Variation of Student Proofs

Visual maps of student proof efforts serve to shape new questions about the how and why of student thinking. In addition, graphic representation provides a snapshot of the range of variation in student efforts. A greater range of variation increases the challenge of providing individualized help.

We use Markov Decision Processes to visualize student proof attempts, as in [4]. A Markov decision process (MDP) is defined by its state set S, action set A, transition probabilities P, and a reward function R [10]. On executing action a in state s the

**Table 4.** Number of low-performing students identified by JT performance levels (Trial Set).

| | *JT1 Success Rate < 90%* | *JT2 Success Rate < 70%* | *JT3 Success Rate < 75%* | *JT1, 2, & 3* | *JT2 & JT3* |
|---|---|---|---|---|---|
| Correctly Categorized | 13 | 17 | 15 | 12 | 15 |
| False Negatives | 4 | 1 | 1 | 1 | 1 |
| False Positives | 3 | 1 | 5 | 7 | 4 |

probability of transitioning to state s′ is denoted P(s′ | s, a) and the expected reward associated with that transition is denoted R(s′ | s, a). For a particular point in a student's proof, our method takes the current premises and the conclusion as the state, and the student's input as the action. Therefore, each proof attempt can be seen as a graph with a sequence of states (each describing the solution up to the current point), connected by actions. We combine all student solution graphs into a single graph, by taking the union of all states and actions, and mapping identical states to one another.

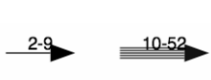
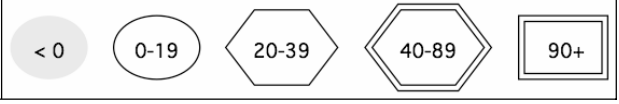
Once this graph is constructed, it represents all of the paths students have taken in working a proof. Typically, a reinforcement learning technique such as value iteration Bellman backup [10] is used to assign reward values to all states in the MDP. The rewards for each state indicate how close to the goal a state is, while probabilities of each transition reveal the frequency of taking a certain action in a certain state. Once the rewards are calculated, the "best" solution to the problem corresponds to taking a path through the graph, choosing nodes of maximum reward to reach the goal state [10]. For this paper, just one step of value iteration was performed to cascade goal values through the MDP.

This experiment uses data from the Spring 2005 and Fall 2005 semesters of the first author's Deductive Logic course, with a total of 69 students. After lectures on the topic and practice using JT, students use DT to solve 12 logic proofs as homework. We extracted 69 students' attempts at solutions to DT proof 1.3. Of these 69 attempts, 52 (75%) were successful proofs, while 17 were incomplete (25%). After cleaning the data, we load the proofs into a database and build a graph for the data. We then set a large reward for the goal state (100) and penalties for incorrect states (10) and a cost for taking each action (1). Setting a non-zero cost on actions penalizes longer solutions (but we set this at 1/10 the cost of taking an incorrect step). These values may need to be adjusted for different problems.

We then created an MDP as described above for the proof data, resulting in a set of 193 states and associated reward values, and 272 actions. Using Excel®, we assigned labels to each state in the MDP (using the latest premises added), colors for errors, state values, and action frequencies, and prepared the data for display. We used GraphViz to display the output. Table 5 shows the legend for nodes and edges. After graphing the MDP, we continually refined the visualization to explore questions about the data.

Figure 1 shows the MDP restricted to the actions that at least 4 students have taken (since the overall MDP is too large to view here). Actions are labeled with a B for backwards steps and F for forward steps, followed by a dash and the rule used. The starting state of the problem contains 3 given premises: (~AvB)>(C&D),A>E, ~E, and the result: ?D. In DT, the premises to be proven are denoted with ?, and when a student works backwards, the original result is shown to be justified by the intermediate

**Table 5.** Legend for MDP edges and nodes

| Edges (Values=Frequency) | Nodes (Values=Rewards) |
|---|---|
|   |   |

expressions the student postulates, and the intermediate expressions become new provisional postulates. For example, many students work backward from ?D using simplification (B-SIMP) to postulate the intermediate expression ?C&D. As shown at the top of Figure 1, many students take this as a first step. In this step, the ? is removed from D and remains on ?C&D until it is justified.

In Figure 1, we see that there are a number of ways students solve problem 1.3. The leftmost part of Figure 1 shows the students who worked primarily backward in solving 1.3, while the central area shows students with forward and mixed approaches. The rightmost part shows an abandoned attempt to work backwards via MP. From this diagram, we note that students working primarily backward seem to have much more trouble determining that they must prove ~A than those students who started using a forward approach. We are planning to build a tool that enables instructors to interactively view student solutions and understand student thinking and approaches to problems.

Figure 2 shows the most frequent forwards actions, taken by at least 9 students each. Sixteen students take the first F-MT action, and nine (56%) of these take F-ADD and achieve the goal (taking the remaining steps shown) through forward-only actions. The 7 other students taking F-MT as a first step used a mixed approach to the problem.

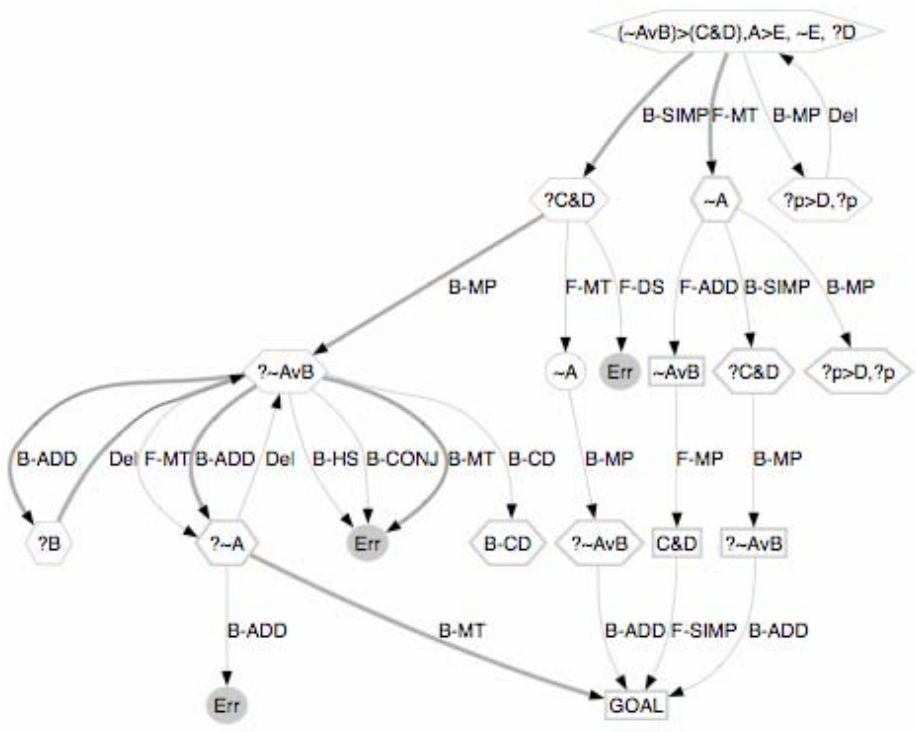



**Figure 1.** MDP of DT1.3 restricted to actions taken by at least 4 students.

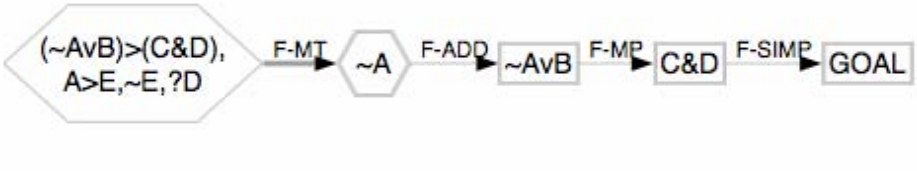



**Figure 2.** MDP restricted to frequent Forward actions, with at least 9 students taking each action.

Figure 3 shows the most frequent backwards actions, taken by at least 17 students each. Fifty-two students take the first B-SIMP action, and 45 of these take B-MP to derive ?~AvB. Then, about half (24) of the students derive ?B, which cannot be proven, and so delete this node and try again. A third of students (17) try using B-MT in error. Eventually, 41 students correctly derive ?~A and 35 of these reach the goal using backwards-only actions. Many of the remaining 17 students eventually succeed using a mixed approach, but this analysis shows a strong preference of students toward using backward-only approaches on this problem. This most likely reflects a focus on this approach in lectures, and also reflects the relative complexity of the premises in this problem. Although there are only 4 main actions that must be taken to solve this proof, we see that students do search to find correct approaches.

In looking at both backwards-only and forwards-only approaches, we see the most divergence for students in the steps connecting ~A with ~AvB, using the addition rule (ADD). This suggests that students do not naturally think of applying addition. The first two authors have observed this anecdotally, but our visualization confirms this reluctance, and demonstrates that students attempt a number of other approaches before trying addition. Based on these findings, we can suggest more discussion of the addition rule in class, and can also plan to build hints into DT when addition applies. In the next section, we discuss how we plan to use the data generated by MDPs and the associated visualizations to target help to individual students using DT.

## 4. Generating Individualized Help

Assuming early identification of students at risk for failing the course, and more particularly, at risk for failing to master proof construction, the question becomes how

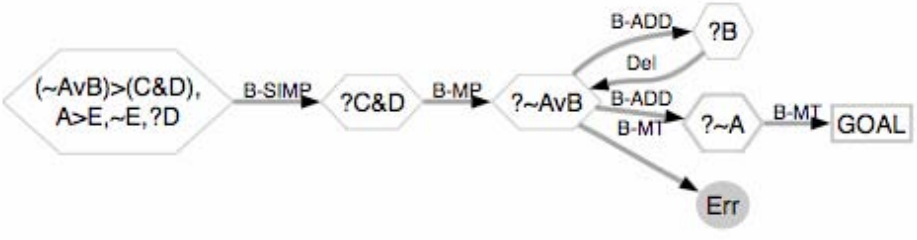



**Figure 3.** MDP restricted to frequent backwards actions, with at least 17 students taking each action.

to effectively address their needs. One alternative is to provide real-time, individualized hints to support on-going student proof construction efforts.

As we have proposed in [4], we plan to generate an MDP for each problem in DT and use it to generate hints for new students solving proofs in DT. Since DT has been used as a computer-aided instructional tool for a number of years, we have many semesters of data from which to create large MDPs for each DT problem. We plan to first use these MDPs to add intelligent hints to every problem. As a new student works a DT problem, we will match their states to those in the MDP. If their state is present in the problem's MDP, we will enable a Hint Button to give contextual help.

In [4], we have proposed several reward functions that could be used in hint generation. The three types of reward functions we have proposed are: 1) expert, 2) typical, and 3) least error-prone. The reward function we have described herein reflects an expert reward function, where the value for a state reflects the shortest path to the goal state. On the other hand, a "typical" reward function will choose a path through the MDP that reflects frequent student responses, giving high rewards to correct responses given by many students. The "least error-prone" function would assign high rewards to paths with low probabilities of errors. We plan to implement MDPs with each of these three rewards calculated for each state.

Given the current state, when the Hint Button is pressed, we will select a reward function for the current student based on their student profile. If we have identified the student as an at-risk student as described in section 3, we may select the "least error-prone" reward function for generating hints. On the other hand, high-performing students would likely benefit from expert hints, while students between these two extremes may benefit from hints reflecting typical student behavior [4].

After we've selected a reward function, we select the next state with the highest reward value. We propose three levels of hints from this state:

1. Tell the student what rule to apply next (rule hint).
2. Indicate the premises where the rule can be used (pointing hint).
3. Tell the student both the rule and the premises to combine, resulting in a "bottom-out" hint (e.g., giving the answer) [2].

We also propose to add a limit on the number of hints a student can use and still receive credit for working the problem. We believe that three hints is a fair amount, to be used on a single state in sequence as above or on separate states in the same problem. This results in giving the student one full step of the proof, or allowing rule hints up to three times.

If a student's state is not found in the MDP, the Hint Button will be disabled. Such a student can get DT's built-in feedback that indicates the correctness of each step, but will not get strategic help. However, we can add the student's action and its correctness to our database, and periodically run reinforcement learning to update the reward function values. Before an update is applied in DT, we will test the update to be sure that the instructors agree with the generated hints.

Once we have tested the feasibility of MDP-generated hints, we may group students according to their DT behavior and class performance, and run MDPs for each group of students. Then when the student asked for hints, the MDP chosen for that student will be tailored to both the current context (problem state), and characteristics of the student. For example, we foresee grouping students as those preferring to use backwards-only actions, forward-only, and mixed approaches, and those at-risk or not. Then, the suggested hints will be more likely to be usable by each student.

## 5. Conclusions and Directions for Future Research

We have proposed a two-pronged approach to using data to improve deductive logic education: 1) using prior course data to find early indicators of poor performance, and 2) deriving a way to leverage past student work in generating individualized help in writing proofs. In the first approach, we have determined that one applet (JT) is particularly important for student success in our Deductive Logic Course. Individualized feedback, reminding students to complete JT with at least 70% success, is one way to improve overall course performance. In the second approach, we have explored visualizations of student solutions to a logic proof in DT to determine other places for individualized help. We have concluded that even in a simple proof problem, there is a need for individualized help.

We have proposed an approach to generating these hints using both student characteristics and prior data. DT can already provide feedback on many errors students make. Adding MDPs to this tutor will enable it to provide individualized hints. These MDPs can constantly learn from new student data. We note that on cold start for a new problem that has no student data, the system will still act as a problem-solving environment, but after even one semester of data is collected, a limited amount of hints can be generated. As more data are added, more automated assistance can be generated. Once implemented, we will test the hints generated based on MDPs. We will investigate the effectiveness of 1) hints tailored according to a student's JT performance and general proof approaches, and 2) hints derived from expert, typical, and least error-prone MDPs. These resources promise to serve as a solid foundation for building our intelligent tutoring system.

## References

[1] J. Hartley and D. Sleeman, Towards more intelligent teaching systems. *International Journal of Man-Machine Studies* 2 (1973), 215-236.

[2] K. VanLehn. The behavior of tutoring systems, *International Journal of Artificial Intelligence in Education* 16 (2006), 227-265.

[3] T. Murray, Authoring intelligent tutoring systems: An analysis of the state of the art, *International Journal of Artificial Intelligence in Education* 10 (1999), 98-129.

[4] T. Barnes and J. Stamper, Toward the extraction of production rules for solving logic proofs, *Proc. 13th Intl. Conf. on Artificial Intelligence in Education, Educational Data Mining Workshop (AIED2007),* Marina del Rey, CA, July 9, 2007.

[5] R.E. Mayer, *Thinking, Problem Solving, Cognition*, Freeman, New York, 1992.

[6] M.J. Croy, Graphic interface design and deductive proof construction, *Journal of Computers in Mathematics and Science Teaching* 18 (1999), 371-386.

[7] M.J. Croy, Problem solving, working backwards, and graphic proof representation, *Teaching Philosophy* 23 (2000), 169-187.

[8] R. Scheines and W. Sieg, Computer environments for proof construction, *Interactive Learning Environments* 4 (1994), 159-169.

[9] E. F. Schumacher, *Small is Beautiful*, Harper and Row, New York, 1973, 89.

[10] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*


# Toward Aligning Computer Programming with Clear Thinking via the Reason Programming Language

Selmer BRINGSJORD[1] & Jinrong LI
*Department of Cognitive Science*
*Department of Computer Science*
*Rensselaer AI & Reasoning Laboratory*
*Rensselaer Polytechnic Institute (RPI)*

**Abstract.** Logic has long set itself the task of helping humans think clearly. Certain computer programming languages, most prominently the Logo language, have been billed as helping young people become clearer thinkers. It is somewhat doubtful that such languages can succeed in this regard, but at any rate it seems sensible to explore an approach to programming that guarantees an intimate link between the thinking required to program and the kind of clear thinking that logic has historically sought to cultivate. Accordingly, Bringsjord has invented a new computer programming language, Reason, one firmly based in the declarative programming paradigm, and specifically aligned with the core skills constituting clear thinking. Reason thus offers the intimate link in question to all who would genuinely use it.

**Keywords.** Logic, thinking, Computer Programming language, Reason, Logo

## Introduction

Logic has long seen itself as the field seeking to foster clear thinking in human persons. This can of course be readily confirmed by consulting any comprehensive introduction to logic, which invariably offers rationales such as that by the study of logic one will be less likely to be hoodwinked by fallacious reasoning.[2] Put in terms of a key distinction that goes back to Aristotle, logic has been viewed as a prescriptive, rather than a descriptive enterprise: a field charged with explaining, in detail, what representations, (and processes over these representations,) ought to be followed by those aspiring to be

[1] Corresponding Author: Selmer Bringsjord, Troy NY 12180 USA, 122007 1400 NY USA; E-mail: selmer@rpi.edu Website: http://www.rpi.edu/~brings

[2] See, e.g., the extensive discussion of fallacies in ([33]). See also ([34]), which has the added benefit of setting out logic in a way designed to reveal it's connection to computer science — a connection central to Reason.

clear thinkers.[3] For people with this ambition, logic enables declarative, (specifically propositional) content to be expressed in syntactically and semantically rigorous ways, at which point the field can then also specify methods of reasoning to be applied to this formalized content, in order to allow agents to know more, to know *why* he or she knows more, and to be able to share this knowledge with others.

Unfortunately, only a small fraction of people study logic; this is true even if the population in question is restricted to the civilized world. With few exceptions across the globe, pre-college mathematics curricula avoid logic and the traditional introduction to formal logic is first available, 99 times out of 100, to first-year college students as an elective. Even those majoring in mathematics or philosophy can often obtain degrees without having to take a course in formal logic. On the other hand, at least across the technological world, computer programming *is* quite often introduced to young students.[4] Furthermore, certain computer programming languages, most prominently the Logo language, have long been billed as helping young people become clearer thinkers. It is however doubtful that such languages can succeed in this regard (for reasons to be briefly discussed below), but as such, it is sensible to explore an approach to programming that *guarantee*s an intimate link between the thinking required to program, and the kind of clear thinking that logic has historically sought to cultivate. Accordingly, Bringsjord has invented a new computer programming language standing at the nexus of computing and philosophy: Reason, a language firmly grounded in the logic-based programming paradigm, and thus one offering the intimate link in question to all who would genuinely use it.

The plan of this paper is as follows. The next section (1) is a barbarically quick introduction to elementary logic, given to ensure that the present paper will be understandable to readers from fields other than logic and philosophy. In section 2, we define what we mean by 'clear thinking,' and provide two so-called *logical illusion*s ([1,2]): that is, two examples of difficult problems which those capable of such thinking should be able to answer correctly (at least after they have Reason at their disposal). The next section (3) is devoted to a brief discussion of Logo, and Prolog and logic programming. Logo and Prolog are essential to understanding the motivation for creating Reason. In section 4, Reason itself is introduced in action, as it's used to solve the two problems posed in section 2. The paper ends with a brief section on the future of Reason.

## 1. Elementary Logic in a Nutshell

In logic, declarative statements, or propositions, are represented by formulas in one or more logical systems, and these systems provide precise machinery for carrying out reasoning. The simplest logical systems that have provided sufficient raw material for building corresponding programming languages are the propositional calculus, and the predicate calculus (or first-order logic, or just FOL); together, this pair comprises what is

[3] Aristotle's work on logic, devoted to setting out the theory of the syllogism, can be found in his Organon, the collection of his logical treatises. This collection, and Aristotle's other main writings, are available in ([35]). For a nice discussion of how Aristotelean logic is relevant to modern-day logic-based artificial intelligence, see ([36]).

[4] In the United States, there is an Advanced Placement exam available to any high school student seeking college credit for demonstrated competence in (Java) programming.

generally called elementary logic. We proceed now to give a very short review of how knowledge is represented and reasoned over in these logical systems.

In the case of both of these systems, and indeed in general when it comes to any logic, three main components are required: one is purely syntactic, one is semantic, and one is metatheoretical in nature. The syntactic component includes specification of the alphabet of a given logical system, the grammar for building well-formed formulas (wffs) from this alphabet, and, more importantly, a proof theory that precisely describes how and when one formula can be inferred from a set of formulas. The semantic component includes a precise account of the conditions under which a formula in a given system is true or false. The metatheoretical component includes theorems, conjectures, and hypotheses concerning the syntactic component, the semantic component, and connections between them. In this paper, we focus on the syntactic side of things. Thorough but refreshingly economical coverage of the formal semantics and metatheory of elementary logic can be found in ([3]).

As to the alphabet for propositional logic, it's simply an infinite list

$$p_1, p_2, \ldots, p_n, p_{n+1}, \ldots$$

of propositional variables (according to tradition $p_1$ is $p$, $p_2$ is $q$, and $p_3$ is $r$), and the five familiar truth-functional connectives $\neg$, $\rightarrow$, $\leftrightarrow$, $\wedge$, $\vee$. The connectives can at least provisionally be read, respectively, as 'not,' 'implies' (or 'if then '), 'if and only if,' 'and,' and 'or.' Given this alphabet, we can construct formulas that carry a considerable amount of information. For example, to say that 'if Alvin loves Bill, then Bill loves Alvin, and so does Katherine' we could write

$$a_1 \rightarrow (b_1 \wedge k_1)$$

where the propositional variables denote the individuals involved in the obvious way. These propositional variables, as you can see, are each used to represent declarative statements.

We move up to first-order logic when we allow the quantifiers $\exists x$ ('there exists at least one thing $x$ such that ...') and $\forall x$ ('for all $x$ ...'); the first is known as the *existential* quantifier, and the second as the *universal*. We also allow a supply of variables, constants, relations, and function symbols. As an example, consider that this machinery allows us to symbolize the proposition 'Everyone loves anyone who loves someone' as

$$\forall x\ \forall y(\exists z Loves(y, z) \rightarrow Loves(x, y))$$

But how does one go about reasoning over these sorts of formulas? This question is answered below when we begin to turn to programs that anticipate those in Reason, and then to Reason programs themselves. In both cases, computer programs are fundamentally chains of deductive reasoning.

But what about a careful account of the *meaning* of formulae in the propositional and predicate calculi?

The precise meaning of the five truth-functional connectives of the propositional calculus is given via truth-tables, which tell us what the value of a statement is, given the truth-values of its components. The simplest truth-table is that for negation, which

informs us, unsurprisingly, that if φ is T (= TRUE) then ¬φ is F (= FALSE; see first row below double lines), and if φ is F then ¬ φ is T (second row).

| $\phi$ | $\neg\phi$ |
|---|---|
| T | F |
| F | T |

Here are the remaining truth-tables.

| $\phi$ | $\psi$ | $\phi \wedge \psi$ |
|---|---|---|
| T | T | T |
| T | F | F |
| F | T | F |
| F | F | F |

| $\phi$ | $\psi$ | $\phi \vee \psi$ |
|---|---|---|
| T | T | T |
| T | F | T |
| F | T | T |
| F | F | F |

| $\phi$ | $\psi$ | $\phi \rightarrow \psi$ |
|---|---|---|
| T | T | T |
| T | F | F |
| F | T | T |
| F | F | T |

| $\phi$ | $\psi$ | $\phi \leftrightarrow \psi$ |
|---|---|---|
| T | T | T |
| T | F | F |
| F | T | F |
| F | F | T |

Notice that the truth-table for disjunction says that when both disjuncts are true, the entire disjunction is true. This is called *inclusive* disjunction. In *exclusive* disjunction, it's one disjunct or another, but not both. This distinction becomes particularly important if one is attempting to symbolize parts of English (or any other *natural language*). It would not do to represent the sentence

George will either win or lose.

as

$$W \vee L,$$

because under the English meaning there is no way both possibilities can be true, whereas by the meaning of $\vee$ it would be possible that W and L are *both* true. One could use $\vee_x$ to denote *exclusive disjunction*, which can be defined through the following truth-table.

| $\phi$ | $\psi$ | $\phi \vee_x \psi$ |
|---|---|---|
| T | T | F |
| T | F | T |
| F | T | T |
| F | F | F |

Given a truth-value assignment v (i.e., an assignment of T or F to each propositional variable P), one can say that v "makes true" or "models" or "satisfies" a given formula $\phi$; this is standardly written

$$v \vDash \phi.$$

A formula such that there is some model that satisfies it is said to be *satisfiable*. A formula that cannot be true on any model (e.g., $p \wedge \neg p$) is said to be *unsatisfiable*. Some formulas are true on all models. For example, the formula $((p \vee q) \wedge \neg q) \rightarrow p$ is in this category. Such formulas are said to be *valid* and are sometimes referred to as *validities*. To indicate that a formula $\phi$ is valid we write

$$\models \phi.$$

Another important semantic notion is *consequence*. An individual formula $\phi$ is said to be a consequence of a set $\Phi$ of formulas provided that all the truth-value assignments on which all of $\Phi$ are true is also one on which $\phi$ is true; this is customarily written

$$\Phi \models \phi.$$

The final concept in the semantic component of the propositional calculus is the concept of consistency: we say that a set $\Phi$ of formulas is *semantically consistent* if and only if there is a truth-value assignment on which all of $\Phi$ are true. As a check of understanding, the reader may want to satisfy herself that the conjunction of formulas taken from a semantically consistent set must be satisfiable.

And now, what about the semantic side of first-order logic?

Unfortunately, the formal semantics of FOL get quite a bit trickier than the truth table-based scheme sufficient for the propositional level. The central concept is that in FOL, formulas are said to be true (or false) on *models*; that some formula $\varphi$ is true on a model is often written as $\mathfrak{I} \models \phi$. (This is often read, "$\mathfrak{I}$ *satisfies*, or models, $\phi$.") For example, the formula $\forall x \exists y Gyx$ might mean, on the standard model for arithmetic, that for every natural number $n$, there is a natural number m such that $m > n$. In this case, the *domain* is the set of natural numbers, that is, $\mathbf{N}$; and G symbolizes 'greater than.' Much more could, of course, be said about the formal semantics (or *model theory*) for FOL — but this is an advanced topic beyond the scope of the present, brief treatment. For a fuller but still-succinct discussion using the traditional notation of model theory see ([3]).

## 2. What is Clear Thinking?

The concept of clear thinking, at least to a significant degree, can be operationally defined with help from psychology of reasoning; specifically with help from, first, a distinction between two empirically confirmed modes of reasoning: context-dependent reasoning, versus context-*in*dependent reasoning; and, second, from a particular class of stimuli used in experiments to show that exceedingly few people can engage in the latter mode. The class of stimuli is what has been called *logical illusions*. We now proceed to explain the distinction and the class.

### *2.1 Context-dependent v. context-independentt reasoning*

In a wide-ranging paper in *Behavioral and Brain Sciences* that draws upon empirical data accumulated over more than half a century, Stanovich & West ([4]) explain that there are two dichotomous systems for thinking at play in the human mind: what they

call System 1 and System 2. Reasoning performed on the basis of System 1 thinking is bound to concrete contexts and is prone to error; reasoning on the basis of System 2 cognition "abstracts complex situations into canonical representations that are stripped of context" ([4], p. 662), and when such reasoning is mastered, the human is armed with powerful techniques that can be used to handle the increasingly abstract challenges of the modern, symbol-driven marketplace. System 1 reasoning is context-dependent, and System 2 reasoning is context-independent. We now explain the difference in more detail.

Psychologists have devised many tasks to illuminate the distinction between these two modes of reasoning (without always realizing, it must be granted, that that was what they were doing). One such problem is the Wason Selection Task ([5]), which runs as follows. Suppose that you are dealt four cards out of a larger deck, where each card in the deck has a digit from 1 to 9 on one side, and a capital Roman letter on the other. Here is what appears to you when the four cards are dealt out on a table in front of you:

| E | K | 4 | 7 |
|---|---|---|---|

Now, your task is to pick just the card or cards you would turn over to try your best at determining whether the following rule is true:

($R_1$) If a card has a vowel on one side, then it has an even number on the other side.

Less than 5% of the educated adult population can solve this problem (but, predictably, trained mathematicians and logicians are rarely fooled). This result has been repeatedly replicated over the past 15 years, with subjects ranging from 7th grade students to illustrious members of the Academy; see ([6]). About 30% of subjects do turn over the E card, but that isn't enough: the 7 card must be turned over as well. The reason is as follows. The rule in question is a so-called *conditional* in formal logic, that is, a proposition having an if-then form, which is often symbolized as $\phi \rightarrow \psi$, where the Greek letters here are variables ranging over formulas from some logical system. As the truth-tables routinely taught to young pre-12 math students make clear (e.g., see Chapter 1 of Bumby, Klutch, Collins & Egbers[7]), a conditional is false if and only if its antecedent, $\phi$, is true, while its consequent, $\psi$, is false; it's true in the remaining three permutations. So, if the E card has an odd number on the other side, ($R_1$) is overthrown. However, if the 7 card has a vowel on the other side, this too would be a case sufficient to refute ($R_1$). The other cards are entirely irrelevant, and flipping them serves no purpose whatsoever, and is thus profligate.

This is the abstract, context-independent version of the task. But now let's see what happens when some context-*de*pendent reasoning is triggered, for there is incontrovertible evidence that *if the task in question is concretized*, System 1 reasoning can get the job done ([8]). For example, suppose one changes rule ($R_1$) to this rule:

($R_2$) If an envelope is sealed for mailing, it must carry a 20 cent stamp on it.

And now suppose one presents four envelopes to you (keeping in mind that these envelopes, like our cards, have a front and back, only one side of which will be visible if the envelopes are "dealt" out onto a table in front of you), viz.,

| sealed envelope | unsealed envelope | env. w/ 20 cent stamp | env. w/ 15 cent stamp |
|---|---|---|---|

Suppose as well that you are told something analogous to what subjects were told in the abstract version of the task, namely, that they should turn over just those envelopes

needed to check whether (R2) is being followed. Suddenly the results are quite different: Most subjects choose the sealed envelope (to see if it has a 20 cent stamp on the other side), *and* this time they choose the envelope with the 15 cent stamp (to see if it is sealed for mailing).

### *2.2 The king-ace problem*

Now we come to a logical illusion, the King-Ace Problem. As we present the problem, it's a slight variant[5] of a puzzle introduced by Johnson-Laird ([9]). Here it is:

> Assume that the following is true:
> 'If there is a king in the hand, then there is an ace in the hand,' or 'If there is not a king in the hand, then there is an ace in the hand,' — but not both of these if-thens are true.
> What can you infer from this assumption? Please provide a careful justification for your answer.

You are encouraged to record your own answer. We return to this problem later, when using Reason to solve it. But please note that the correct answer to the problem is not 'There is an ace in the hand,' but rather the (counterintuitive!) proposition that there is *not* an ace in the hand. If Reason is on the right track, use of it will help students see that this is the right answer.

### *2.3 The wine drinker problem*

Now let us consider a second logical illusion, an interesting puzzle devised by Johnson-Laird & Savary ([10]) that has the same general form as Aristotle's syllogisms:

> Suppose:
>
> - All the Frenchmen in the restaurant are gourmets.
> - Some of the gourmets are wine drinkers.
>
> Does it follow that some of the Frenchmen are wine drinkers? Please provide a careful justification for your answer.

We will return to this problem later, when using Reason to solve it. But note for now that the correct answer is 'No.' Reason will itself provide a justification for this negative answer.

## 3. The Logo Programming Language; Logic Programming

When youth learn to program by using Logo,[6] by far the programming language most used in the United States to teach programming in grades 6–12, almost without exception, they produce instructions designed to drive a turtle through some sequence of states. For example, the procedure

[5] The variation arises from disambiguating Johnson-Laird's `s or else s0' as `either s or s0, but not both.'

[6] http://el.media.mit.eduLogo-foundation/logo/programming.html

```
to square
repeat 4 [forward 50 right 90]
end
```

causes the turtle to draw a square. In a second, more sophisticated mode of programming, the Logo programmer can process lists in ways generally similar to those available to the Lisp programmer. Ever since a seminal paper by Black, Swan & Schwartz ([11]), it has been known that while students who program in the second way do seem to thereby develop some clearer thinking skills, the improvement is quite slight, the cognitive distance from processing lists to better logical reasoning is great, and hence *transfer* from the first activity to the second is very problematic.[7] In an intelligent reaction to this transfer challenge, Black et al. make a move that is quite interesting from the perspective of our own objective, and the language Bringsjord has built to meet it: viz., they consider whether teaching Prolog[8] might be a better strategy for cultivating in those who learn it a significant gain in clear thinking. Unfortunately, there are five fatal problems plaguing the narrow logic programming paradigm of which Prolog is a concretization. Here's the quintet, each member of which, as shall soon be seen, is overcome by Reason:

1. Logic programming is based on a fragment of full first-order logic: it's inexpressive. Human reasoning, as is well-known, not only encompasses full first-order logic (and hence on this score alone exceeds Prolog), but also modal logic, deontic logic, and so on.
2. Logic programming is "lazy." By this we mean that the programmer doesn't herself construct an argument or proof; nor for that matter does she create a model or countermodel.
3. While you can issue queries in Prolog, all you can get back are assignments to variables, not the proofs that justify these assignments.
4. In addition, Prolog can't return models or counter-models.
5. Finally, as to deductive reasoning, Prolog locks those who program in it into the rule of inference known as *resolution.*[9] Resolution is not used by humans in the business of carrying out clear deductive thinking. Logic and mathematics, instead, are carried out in what is called *natural deduction* (which is why this is the form of deduction almost invariably taught in philosophy and mathematics, two prominent fields among those directly associated with the cultivation of clear thinking in students).

## 4. The Reason Programming Language

### *4.1. Reason in the context of the four paradigms of computer programming*

There are four programming paradigms: *procedural*, reflected, e.g., in Turing machines themselves, and in various "minimalist" languages like those seen in foundational computer science texts (e.g., Davis & Weyuker [12], Pascal, etc.); *functional*, reflected, e.g., in Scheme, ML, and purely functional Common Lisp ([13,14]) *object-oriented,*

[7] In particular, it turns out that making a transition from \plug-and-chug" mathematics to being able to produce proofs is a very difficult one for students to achieve. See ( [37]). For further negative data in the case of Logo, see e.g., ([38], [39]).

[8] There is of course insufficient space to provide a tutorial on Prolog. We assume readers to be familiar with at least the fundamentals. A classic introduction to Prolog is ([18]).

[9] All of resolution can essentially be collapsed into the one rule that from $p \vee q$ and $\neg p$ one can infer q.

reflected, e.g., in Simula, Smalltalk, and Java ([15],[16], [17]); and *declarative*, reflected, albeit weakly, in Prolog ([18]). Reason is in, but is an extension of, the declarative paradigm.[10]

Reason programs are specifically extensions and generalizations of the long-established concept of a *logic program* in computer science (succinctly presented, e.g., in the chapter "Logic Programming" in [3]).

### *4.2. Proofs as programs through a simple denotational proof language*

In order to introduce Reason itself, we first introduce the syntax within it currently used to allow the programmer to build purported proofs, and to then evaluate these proofs to see if they produce the output (i.e., the desired theorem). This syntax is based on the easy-to-understand type-α denotational proof language NDL invented by Konstantine Arkoudas (for background see [19],[20]) that corresponds for the most part to systems of Fitch-style natural deduction often taught in logic and philosophy. Fitch-style natural deduction was first presented in 1934 by two thinkers working independently to offer a format designed to capture human mathematical reasoning as it was and is expressed by real human beings: Gentzen ([21]) and Jaskowski ([22]). Streamlining of the formalism was carried out by Fitch ([23]). The hallmark of this sort of deduction is that assumptions are made (and then discharged) in order to allow reasoning of the sort that human reasoners engage in.

Now here is a simple deduction in NDL, commented to make it easy to follow. This deduction, upon evaluation, produces a theorem that Newell and Simon's Logic Theorist, to great fanfare (because here was a machine doing what "smart" humans did), was able to muster at the dawn of AI in 1956, at the original Dartmouth AI conference.

```
// Here is the theorem to be proved,
// Logic Theorist's "claim to fame":
// (p ==> q) ==>      (~q ==> ~p)

Relations p:0, q:0.   // Here we declare that we have two
                                  // propositional variables, p and q.
                                  // They are defined as 0-ary relations.

// Now for the argument. First, the antecedent (p ==> q)
// is assumed, and then, for contradiction, the antecedent
// (~q) of the consequent (~q ==> ~p).
assume p ==> q
      assume ~q
            suppose-absurd p
                        begin
                           modus-ponens p ==> q, p;
                           absurd q, ~q
                        end
```

[10] As is well known, in theory any Turing-computable function can be implemented through code written in any Turing-complete programming language. There is nothing in principle precluding the possibility of writing a program in assembly language that, at a higher level of abstraction, processes information in accordance with inference in many of the logical systems that Reason allows its programmers to work in. (In fact, as is well-known, the other direction is routine, as it occurs when a high-level computer program in, say, Prolog, is compiled to produce code corresponding to low-level code; assembly language, for example.) However, the mindset of a programmer working in some particular programming language that falls into one of the four paradigms is clearly the focus of the present discussion, and Turing-completeness can safely be left aside.

If, upon evaluation, the desired theorem is produced, the program is successful. In the present case, sure enough, after the code is evaluated, one receives this back:

Theorem: (p ==>q) ==> (~q ==> ~p)

Now let us move up to programs written in first-order logic, by introducing quantification. As you will recall, this entails that we now have at our disposal the quantifiers $\exists x$ ('there exists at least one thing $x$ such that ...') and $\forall x$ ('for all $x$ ...'). In addition, there is now a supply of variables, constants, relations, and function symbols; these were discussed above. What follows is a simple NDL deduction at the level of first-order logic that illuminates a number of the concepts introduced to this point. The code in this case, upon evaluation, yields the theorem that Tom loves Mary, given certain helpful information. It is important to note that both the answer and the justification have been assembled, and that the justification, since it is natural deduction, corresponds to the kinds of arguments often given by human beings.

```
Constants mary, tom. // Two constants announced.

Relations Loves:2.  // This concludes the simple signature, which
                    // here declares Loves to be a two-place relation.

// That Mary loves Tom is asserted:
assert Loves(mary, tom).

// 'Loves' is a symmetric relation, and this is asserted:
assert (forall x (forall y (Loves(x, y) ==> Loves(y, x)))).

//Now the evaluable deduction proper can be written:
suppose-absurd ~Loves(tom, mary)
    begin
        specialize (forall x (forall y (Loves(x, y) ==> Loves(y, x)))) with mary;
        specialize (forall y (Loves(mary, y) ==> Loves(y, mary))) with tom;
        Loves(tom,mary) BY modus-ponens
                        Loves(mary, tom) ==> Loves(tom, mary), Loves(mary, tom);
    end;
Loves(tom,mary) BY double-negation ~~Loves(tom,mary)
```

When this program is evaluated, one receives the desired result back: Theorem: Loves(tom,mary). Once again, note that both the answer and the justification have been assembled, and that the justification, since it is natural deduction, corresponds to the kinds of proofs often crafted by human beings.

So far we have conceived of programs as proof-like entities. But what about the semantic side? What about models? Moving beyond NDL, in Reason, programs can be written to produce, and to manipulate, models. In addition, while in NDL the full cognitive burden is borne by the programmer, Reason can be queried about whether certain claims are provable. In addition, in Reason, the programmer can set the degree to which the system is intelligent on a session-by-session basis. This last property of Reason gives rise to the concept that the system can be set to be "oracular" at a certain level. That is, Reason can function as an oracle up to a pre-set limit. One common limit is propositional inference, and the idea here is to allow Reason to be able to prove on its own anything that requires only reasoning at the level of the propositional calculus.

### *4.3. Cracking king-ace and wine drinker with reason*

#### *4.3.1. Cracking the king-ace problem with reason*

In this example, the selected logic to be used with Reason is standard first-order logic as described above, with the specifics that reasoning is deductive and Fitch-style. The system is assumed to have oracular power at the level of propositional reasoning, that is, the programmer can ask Reason itself to prove things as long as only reasoning at the level of the propositional calculus is requested. This request is signified by use of prop.

To save space, we assume that the programmer has made these selections through prior interaction with Reason. Now, given the following two propositions, is there an ace in the hand? Or is it the other way around?

**F1** If there is a king in the hand, then there is an ace in the hand; or: if there isn't a king in the hand, then there is an ace in the hand.

**F2** Not both of the if-thens in F1 are true.

In this case, we want to write a Reason program that produces the correct answer, which is "There is not an ace in the hand." We also want to obtain certification of a proof of this answer as additional output from our program.

We can obtain what we want by first declaring our symbol set, which in this case consists in simply declaring two propositional variables, K (for 'There is a king in the hand') and A (for 'There is an ace in the hand'). Next, in order to establish what is known, we present the facts to Reason. Note that Reason responds by saying that the relevant things are known, and added to a knowledge base (KB1).

```
>(known F1 KB1 (or (if K A) (if (not K) A)))
F1 KNOWN
F1 ADDED TO KB1

>(known F2 KB1
(not (and (if K A) (if (not K) A))))

F2 KNOWN
F2 ADDED TO KB1
```

Next, we present the following partial proof to Reason. (It's a *partial* proof because the system itself is called upon to infer that the negation of a conditional entails a conjunction of the antecedent and the negated consequent. More precisely, from (not (if P Q)) it follows that (and P (not Q)).)

```
(proof P1 KB1
   demorgan F2;
   assume (not (if K A))
      begin
           (and K (not A)) by prop on (not (if K A));
           right-and K, (not A)
       end
   assume (not (if (not K) A))
       begin
```

```
            (and (not K) (not A) by prop on (not (if (not K) A)));
            right-and (not K), (not A)
        end
proof-by-cases (or (not (if K A)) (not (if (not K) A))),
                        (if (not (if K A)) (not A)),
                        (if (not (if (not K) A)) (not A)))
```

When this proof is evaluated, Reason responds with:

```
PROOF P1 VERIFIED
ADDITIONAL KNOWNS ADDED TO KB1:
THEOREM: (not A)
```

We make the perhaps not unreasonable claim that anyone who takes the time to construct and evaluate this program (or for that matter any reader who takes the time to study it carefully to see why (not A) is provable) doesn't succumb to the logical illusion in question any longer. Now we can proceed to issue an additional query:

```
> (provable? A)
NO
DISPLAY-COUNTERMODEL OFF
```

If the flag for countermodeling was on, Reason would display a truth-table showing that A can be false while F1 and F2 are true.

*4.3.2. Cracking the wine drinker problem with reason*

Recall the three relevant statements, in English:

**F3** All the Frenchmen in the restaurant are gourmets.

**F4** Some of the gourmets are wine drinkers.

**F5** Some of the Frenchmen in the restaurant are wine drinkers.

To speed the exposition, let us assume that the Reason programmer has asserted these into knowledgebase KB2, using the expected infix syntax of first-order logic, so that, for example, F3 becomes

(forall x (if (Frenchman x) (Gourmet x)))

In addition, let us suppose that Reason can once again operate in oracular fashion at the level of propositional reasoning, that the flag for displaying countermodels has been activated, and that we have introduced the constants object-1 and object-2 to Reason for this session. Given this, please study the following interaction.

```
> (proof P2 KB2
    begin
       assume (and (Frenchman object-1) (Gourmet object-1) (not (Wine-drinker object-1)));
       assume (and (Gourmet object-2) (In-restaurant object-2) (Wine-drinker object-2));
       not-provable (F3 F4 F5) => (some x (and (In-restaurant x) (Wine-drinker x)))
    end)
PROOF P2 VERIFIED
DISPLAY-COUNTERMODEL?

>Y
```

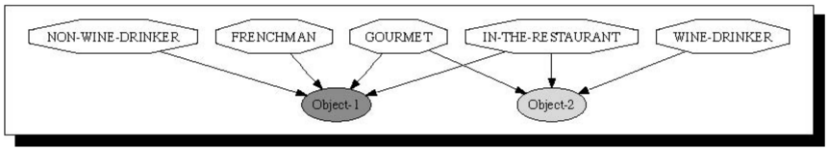



**Figure 1.** Visual Countermodel in Wine Drinker Puzzle (provided via a grammar and program written by Andrew Shilliday and Joshua Taylor that Reason can call)

In the situation the programmer has imagined, all Frenchmen are gourmets, and there exists someone who is a wine-drinker and a gourmet. This ensures that both the first two statements are true. But it's *not* true that there exists someone who is both a Frenchman and a wine drinker. This means that F5 is false; more generally, it means that F5 isn't a deductive consequence of the conjunction of F3 and F4.

When countermodels are rendered in visual form, they can be more quickly grasped. Figure 1 shows such a countermodel relevant to the present case.

## 5. The Future

The full and fully stable implementation of Reason is not yet complete. Fortunately, this implementation is not far off, as it is aided by the fact that this implementation is to a high degree meta-programming over computational building blocks that have been provided by others.[11] For example, resolution-based deduction is computed by the automated theorem provers Vampire [24] and Otter [25,26] (and others as well), while natural deduction-style automated theorem proving is provided by the Oscar system invented by the philosopher John Pollock [27,28]. As to automated model finding for first-order logic, a number of mature, readily available systems now compute this relation as well, for example Paradox and Mace [29]. At the propositional level, truth-value assignments are automatically found by many SAT solvers (e.g., see [30]).

While it seems sensible to strive for teaching clear thinking via programming languages that, by their very nature, are more intimately connected to the formalisms and processes that (from the perspective of logic, anyway) constitute clear thinking, noting this is not sufficient, obviously. One needs to empirically test determinate hypotheses. We need, specifically, to test the hypothesis that students who learn to program in Reason will as a result show themselves to be able, to a higher degree, to solve the kind of problems that are resistant to context-dependent reasoning. Accordingly, empirical studies of the sort we have carried out for other systems (e.g., [31,32]) are being planned for Reason.

## Acknowledgements

Bringsjord is greatly indebted to those who provided insightful comments and objections at E-CAP 2007. More generally, Bringsjord wishes to thank all those who made this conference possible, since the theme driving it (viz., the confluence of computing and philosophy) stands at the heart of Reason. Finally, Bringsjord is grateful to Danny

[11] For a discussion of meta-programming in the logic programming field, see ([32]).

Bobrow, one of the creators of Logo, for conversation about Logo, and about learning to simultaneously program and reason well.

## References

[1] P. N. Johnson-Laird, P. Legrenzi, V. Girotto, and M. S. Legrenzi, "Illusions in reasoning about consistency," *Science*, vol. 288, pp. 531–532, 2000.
[2] S. Bringsjord and Y. Yang, "Logical illusions and the welcome psychologism of logicist artificial intelligence," in *Philosophy, Psychology, and Psychologism: Critical and Historical Essays on the Psychological Turn in Philosophy* (D. Jacquette, ed.), pp. 289–312, Dordrecht, The Netherlands: Kluwer, 2003.
[3] H. D. Ebbinghaus, J. Flum, and W. Thomas, *Mathematical Logic (2nd edition)*. New York, NY: Springer-Verlag, 1994.
[4] K. E. Stanovich and R. F. West, "Individual differences in reasoning: Implications for the rationality debate," *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 645–665, 2000.
[5] P. Wason, "Reasoning," in *New Horizons in Psychology*, Hammondsworth, UK: Penguin, 1966.
[6] S. Bringsjord, E. Bringsjord, and R. Noel, "In defense of logical minds," in *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pp. 173–178, Mahwah, NJ: Lawrence Erlbaum, 1998.
[7] Bumby, Klutch, Collins, and Egbers, *Integrated Mathematics Course 1*. New York, NY: Glencoe/McGraw Hill, 1995.
[8] M. Ashcraft, *Human Memory and Cognition*. New York, NY: HarperCollins, 1994.
[9] P. Johnson-Laird, "Rules and illusions: A criticial study of Rips's *The Psychology of Proof*," *Minds and Machines*, vol. 7, no. 3, pp. 387–407, 1997.
[10] P. Johnson-Laird and F. Savary, "How to make the impossible seem probable," in *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (M. Gaskell and W. Marslen-Wilson, eds.), pp. 381–384, Hillsdale, NJ: Lawrence Erlbaum Associates, 1995.
[11] J. Black, K. Swan, and D. Schwartz, "Developing thinking skills with computers," *Teachers College Record*, vol. 89, no. 3, pp. 384–407, 1988.
[12] M. Davis, R. Sigal, and E. Weyuker, *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*. New York, NY: Academic Press, 1994.
[13] S. Shapiro, *Common Lisp: An Interactive Approach*. New York, NY: W. H. Freeman, 1992.
[14] H. Abelson and G. Sussman, *Structure and Interpretation of Computer Programs (2nd Edition)*. Cambridge, MA: MIT Press, 1996.
[15] K. Nygaard and O.-J. Dahl, "The development of the simula languages," pp. 439–480, 1981.
[16] A. C. Kay, "The early history of smalltalk," pp. 511–598, 1996.
[17] K. Arnold, J. Gosling, and D. Holmes, *Java(TM) Programming Language, The (4th Edition)*. Addison-Wesley Professional, 2005.
[18] W. Clocksin and C. Mellish, *Programming in Prolog (Using the ISO Standard; 5th Edition)*. New York, NY: Springer, 2003.
[19] K. Arkoudas, *Denotational Proof Languages*. PhD thesis, MIT, Department of Computer Science, Cambridge, USA, 2000.
[20] S. Bringsjord, K. Arkoudas, and P. Bello, "Toward a general logicist methodology for engineering ethically correct robots," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 38–44, 2006.
[21] G. Gentzen, "Untersuchungen über das logische Schlieben I," *Mathematische Zeitschrift*, vol. 39, pp. 176–210, 1935.
[22] S. Jaskowski, "On the rules of suppositions in formal logic," *Studia Logica*, vol. 1, 1934.
[23] F. Fitch, *Symbolic Logic: An Introduction*. New York, NY: Ronald Press, 1952.
[24] A. Voronkov, "The anatomy of vampire: Implementing bottom-up procedures with code trees," *Journal of Automated Reasoning*, vol. 15, no. 2, 1995.
[25] L. Wos, *The Automation of Reasoning: An Experimenter's Notebook with OTTER Tutorial*. San Diego, CA: Academic Press, 1996.
[26] L. Wos, R. Overbeek, e. Lusk, and J. Boyle, *Automated Reasoning: Introduction and Applications*. New York, NY: McGraw Hill, 1992.
[27] J. Pollock, *How to Build a Person: A Prolegomenon*. Cambridge, MA: MIT Press, 1989.
[28] J. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press, 1995.
[29] K. Claessen and N. Sorensson, "New techniques that improve Mace-style model finding," in *Model Computation: Principles, Algorithms, Applications (Cade-19 Workshop)*, (Miami, Florida), 2003.

[30] H. Kautz and B. Selman, "Unifying SAT-based and graph-based planning," in *Workshop on Logic-Based Artificial Intelligence, Washington, DC, June 14– 16, 1999* (J. Minker, ed.), (College Park, Maryland), Computer Science Department, University of Maryland, 1999.
[31] K. Rinella, S. Bringsjord, and Y. Yang, "Efficacious logic instruction: People are not irremediably poor deductive reasoners," in *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (J. D. Moore and K. Stenning, eds.), pp. 851–856, Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
[32] S. Bringsjord and D. Ferrucci, "Logic and artificial intelligence: Divorced, still married, separated...?," *Minds and Machines*, vol. 8, pp. 273–308, 1998.
[33] I. Copi and C. Cohen, *Introduction to Logic*. Englewood Cliffs, NJ: Prentice-Hall, 1997. This is the tenth edition of the book.
[34] J. Barwise and J. Etchemendy, *Language, Proof, and Logic*. New York, NY: Seven Bridges, 1999.
[35] R. McKeon, ed., *The Basic Works of Aristotle*. New York, NY: Random House, 1941.
[36] C. Glymour, *Thinking Things Through*. Cambridge, MA: MIT Press, 1992.
[37] R. C. Moore, "Making the transition to formal proof," *Educational Studies in Mathematics*, vol. 27.3, pp. 249–266, 1994.
[38] K. Louden, "Logo as a prelude to lisp: Some surprising results," *ACM SIGCSE Bulletin*, vol. 21, no. 3, 1989.
[39] R. D. Pea, "Logo programming and problem solving: Children's experiences with logic," in *Educational computing (An Open University Reader)* (T. O'Shea and E. Scanlon, eds.), London, UK: John Wiley and Sons, 1987.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Ethics and the Practice of Software Design

Matteo TURILLI [1, 2, 3, 4, 5]

*[1] Oxford University Computing Laboratory (OUCL), [2] Oxford e-Research Centre (OeRC) and [3] Information Ethics Group (IEG), University of Oxford, UK, [4]Research Group in Philosophy of Information (GPI), University of Hertfordshire,*
*[5]Centre for Ethics and Economics and Business, Universidade Católica Portugueasa*

**Abstract.** The paper offers an analysis of the problem of integrating ethical principles into the practice of software design. The approach is grounded on a review of the relevant literature from Computer Ethics and Professional Ethics. The paper is divided into four sections. The first section reviews some key questions that arise when the ethical impact of computational artefacts is analysed. The inner informational nature of such questions is used to argue in favour of the need for a specific branch of ethics called Information Ethics. Such ethics deal with a specific class of ethical problems and Informational Privacy is introduced as a paradigmatic example. The second section analyses the ethical nature of computational artefacts. This section highlights the fact that this nature is impossible to comprehend without first considering designers, users, and patients alongside the artefacts they create, use and are affected by. Some key ethical concepts are discussed, such as freedom, agency, control, autonomy and accountability. The third section illustrates how autonomous computational artefacts are rapidly changing the way in which computation is used and perceived. The description of the ethical challenges posed to software engineers by this shift in perspective closes the section. The fourth and last section of the paper is dedicated to a discussion of Professional Ethics for software engineers. After establishing the limits of the professional codes of practice, it is argued that ethical considerations are best embedded directly into software design practise. In this context, the Value Sensitive Design approach is considered and insight into how this is being integrated into current research in ethical design methodologies is given.

**Keywords.** Computer Ethics, Information Ethics, Requirement Engineering, Autonomy, Control Closure

## Introduction

Design, development and deployment of computational artefacts raise challenges in different research areas. Some of these challenges are related to the computation itself – building more efficient computers, devising new algorithms and programming languages, and studying the mathematical properties of computational theories. Other challenges concern the social implications of such artefacts. The study of the ethical impact of computational artefacts falls into the latter category and is best considered from a multidisciplinary stance, by relying on insights coming from professional ethics, applied ethics, metaethics, philosophy of technology, social sciences and requirement engineering.

The importance of drawing a link between computational challenges *per se* and the ethical impact of computational artefacts is best appreciated when one considers that both elements are essential for developing efficient computational artefacts. Eliciting, representing and verifying functional requirements is as important as taking into consideration non-functional requirements such as reliability, scalability, cost effectiveness and the respect for policies, legislation and ethics. One of the tasks of requirement engineering is to bridge the divide existing between these two classes of requirements. Once all the necessary elements – functional and non-functional – are elicited, they must be translated into the design procedures necessary for the development of the artefacts. The required shift from an informal to a formal approach is what makes the bridging goal of requirement engineering particularly challenging.

This paper seeks to contribute to the understanding of how ethics could and should be integrated into the practice of software design. The overall goal is to enable engineers to control the ethical implications of the computational artefacts they design, develop, deploy and maintain. If properly exercised, this control would minimize the negative ethical impact that a growing number of computational artefacts have on their deployment environment. Alternatively considered, positive ethical implications could be maximised transforming computational artefacts into instruments for applied ethics.

## 1. Computational Artefacts as Extensions of Human Informational Capabilities

Research into the design and development of computational artefacts has often gone hand in hand with the necessity of understanding how to reason about their ethical consequences. Without such understanding there is the risk of creating artefacts detrimental to the environment in which they will be deployed.

It is not surprising then that T. W. Bynum, while investigating the origins of Information Ethics, traces a parallel between the ethical standings of the mathematician Norbert Wiener and Aristotle's flourishing Ethics.[1,2] Wiener not only developed the science of information and feedback systems called cybernetics but also had a clear intuition of the deep ethical consequences of his research. Following such intuition, Wiener laid down many ideas developed today by Computer Ethics, Information Ethics and Ethics of ICT.[3,4]

Bynum opens the analysis of Wiener's ethical positions by underlining the following passages. With reference to "ultra-rapid computing machines" Wiener argues that:

> "Long before Nagasaki and the public awareness of the atomic bomb, it had occurred to me that we were here in the presence of another social potentiality of unheard-of importance for good and for evil." [3]

And

> "The answer, of course, is to have a society based on human values other than buying and selling." [3]

Bynum uses this quote to illustrate Wiener's awareness of the impact of computational artefacts on

> "*fundamental human factors*, such as life and health, work and wealth, knowledge and ability, creativity and happiness, democracy and freedom, peace and security". [5]

Wiener's quote is important to grasp two additional points. First, Computing machines as intended by Wiener are instruments analogous to bombs, i.e., artefacts used by humans to achieve their ethical or unethical goals. This way of viewing computational artefacts restricts the relative ethical interrogatives as to whether and how the artefacts ought to be used and for what final purpose.

Second, the answer to the questions of 'how' and 'to what end' is "to have a society based on human values". Such a goal suggests the development of a specific branch of ethics dedicated to the analysis of the ethical implications of computational artefacts. Such ethics would take into account how individuals are affected by computational artefacts and, depending on their effects, whether and how they should be deployed.

Eventually, this branch of ethics has developed in different ways, depending on the focus of the ethical investigation. 'Computer Ethics' indicates the field of research related to the ethical issues produced, aggravated or transformed by computational technologies and the computing profession. 'Information Ethics' has been adopted to refer to the research concerned with the ethical consequence of information itself, including cases in which information involves computation. 'Cyberethics' and 'Ethics of ICT' have been used for investigations specifically related to the ethical consequences of digital information and ICT (information and communication technologies).

Bynum uses 'Information Ethics' when referring to Wiener's work. This choice reflects the fact that the ethical consequences of Wiener's ultra-rapid computing machines depend on their ability to produce and manipulate information and not on architectural details. In other words, it is not necessary to have an electronic implementation of a computational machine to raise ethical issues. The Hollerith machine is a paradigmatic example of how purely electromechanical artefacts have raised the same ethical problems concerning Wiener far before his cybernetics.

H. Hollerith was a German-American engineer who invented a machine capable of tabulating data read from punched cards. Developed for the operations of the American census of 1890, the Hollerith's tabulator reduced the time needed to complete the task by two thirds. While the 1880 census had required seven years, the 1890 census, with a substantially increased population, was double-checked and concluded in just two and a half years.[6]

The Hollerith machine was capable of performing tabulations at a speed and precision previously unachievable, reducing to tractable proportions logistical problems otherwise unsolvable. In 1933, Hitler understood that the attainment of his goals of ethnic cleansing and political domination could be advanced by the pervasive use of data tabulation. He subscribed to a commercial agreement with IBM and deployed Hollerith machines in almost every sector of the Reich's activities. Data related to railway, census and concentration camps were all recorded and tabulated thanks to millions of punched cards.[7]

The Hollerith machines were used for empowering human capabilities as in the case of weapons, writing machines, hardware and Wiener's feedback systems. The deployment of Hollerith machines resulted in a hybrid, human/machine system capable of managing information in order to achieve ethically questionable goals with renewed speed and efficiency. Such computational artefacts inherited and amplified ethical challenges belonging to any human actions devoted to producing and managing information. It remains an open question whether the nature of the problems raised by computational-based informational activity is so original as to justify the existence of a dedicated branch of ethics.

J. H. Moor focuses on two distinctive characteristics of the modern general-purpose computers for understanding why the nature of the problems investigated by Computer Ethics[1] is original.[8] Moor argues that computers are unique artefacts as they are "logically malleable" and "informationally enriching". They are "logically malleable" because they are general-purpose machines that can be programmed to perform tasks in different semantic domains. They are "informationally enriching" because computers transform the nature of the task they perform, making it intrinsically informational. This enrichment gradually changes human activities once they become computer-based. Paradigmatic examples are the design procedures in engineering, the evolution of monetary systems and the progressive digitalisation of the battlefield. Following Moor, all these areas have now become information-based because of the introduction of dedicated computational artefacts and their informational enriching power.

When considering logical malleability and informational enrichment, Moor rejects the hypothesis for which the problems of Computer Ethics do not have a specific nature, distinguishable from those of other ethical fields. His analysis of the concept of Informational Privacy demonstrates how sharing digitalised data is the main reason privacy has evolved from being concerned about physical intrusions to be focussed on the protection of an individual's data. These data, once digitalised, become "greased" [9] meaning that they can be seamlessly duplicated, transported, stored, mined and correlated. While the problem of Informational Privacy is as old as cameras, [10] it is only with the digitalisation of information that it assumes the contemporary characterisation.

After almost ten years, the challenging practical effects of Moor's analysis have become evident. Identity theft is a macroscopic example of the kind of ethical and legal problems that digitalisation of personal data and violation of Informational Privacy generates. The last report on this crime estimates that in 2007, 8.4 million American

[1] The problems of Computer and Information ethics overlap when computation is involved in the production, flow and management of information. The characteristics of computation may affect information in an ethically meaningful way and, conversely, the informational component of computation can generate ethical concerns. Information Ethics is here used referring directly to computation and, as such, it can be considered part of Computer Ethics.

citizens were affected with damages totalling $56.6 billion.[11] Identity theft is a complex crime [12], made possible mainly by (i) unauthorised access to sensitive data almost always stored in digital form and (ii) the ability to impersonate the fraud victim, which entails being authenticated by the automated systems of identity verification.

The specific informational nature of computational artefacts and how it affects ethical problems redefines their nature thereby justifying the creation of a specific branch of applied ethics. It remains to be seen how these specific problems should be tackled.

Computer Ethics can seek to define the best policies for the deployment of some artefacts but can do much more if oriented towards the process of designing them. In this way, Computer Ethics would not be limited to contributing towards managing the ethical impact of computational artefacts but it would contribute in shaping it at its source.

Following this direction, the next section looks at the artefacts themselves, how their functionalities are designed, their 'soft' nature and how these affect their users and patients.

## 2. Deconstructing Computational Artefacts: Software, Hardware, Users, Designers and the Environments in which They Operate

In the social scenario generated by the mass diffusion of Personal Computers (PCs), designing computational artefacts has started to affect a larger proportion of the population generating and/or amplifying ethical problems. As noted by Moor, ethical problems are transformed in their nature by the informational properties of computers. It is therefore useful to clarify what specific characteristics enable computational artefacts to generate their ethical impact and how the social environment in which these artefacts operate contributes to it.

The distinction between software and hardware helps to clarify that the behaviour of computational artefacts depends on a design. The practise of software design determines the functionalities that software will exhibit once deployed – what information will be considered (input), how it will be managed in order to produce other information (output) – but also the qualities of the software such as, for example, its reliability, usability or computational efficiency.

When combined, the qualitative and functional characteristics of computational artefacts determine their ethical impact. 'Bugs', imperfections in design and erroneous assumptions about usability are all factors that produce mismatches between how computational artefacts behave and the environment in which they are deployed. There are different types of mismatches: reliability (artefacts that do not perform the right operations for one or more situations); cost effectiveness (artefacts require too many resources); usability (artefacts that do not offer a proper interface to their users); but also ethical fitness (artefacts do not respect one or more ethical principles endorsed in their deployment environment).

It is now clear that the control of the ethical impact of computational artefacts can occur at the moment of their 'conception', i.e., at design time. Software engineers are accountable for the mismatches described above, as these are a direct result of their design choices (or lack thereof). It follows that ethics and the practice of software design should be combined in order to avoid undesirable consequences. Unfortunately,

as usual with software design, not only explicit errors but also lack of specification (indeterminism) can lead to the most dramatic consequences.

Between 2000 and 2001, 28 patients of the Panamanian National Oncological Institute were treated for cancer with excessive doses of radiation. This accident caused the death of several patients and serious injuries for others. The erroneous lethal doses had been calculated by a computer-based Treatment Planning System (TPS). In this case, the TPS was used in a way not anticipated by its designers. TPS was designed to take into account the use of up to four shields when calculating the dose of radiation to administer to patients. In this instance, TPS users set out to discover an efficient, cost effective way for using TPS to calculate the radiation doses with *five* shields. Unfortunately, in such conditions the software miscalculated radiation doses. Such calculations went unchecked, leading to erroneous treatment and deleterious consequences. This accident escalated to international proportions with the involvement of the International Atomic Energy Agency (IAEA) and promoted a debate about who should be held accountable for the deaths and injuries and how the negative outcomes could have been avoided.[13]

This example dramatically highlights the problem of accountability but also how the ethical dimension of computational artefacts extends outside the boundaries of the artefacts' to include their designers, users and patients.

In order to clarify the properties of this extended ethical dimension of the artefacts, it is useful to define how computational artefacts relate to core ethical concepts. Examples of these concepts are autonomy, control and agency. Understanding whether computational artefacts can be considered agents, in which sense they express autonomy, and how they differ in performing their operations from humans performing actions, are three preconditions in order to discuss the limits of the accountability of computational artefacts, when and where they should be deployed and, more generally, how they should be designed from an ethical point of view.

D. Johnson assumes an orthodox definition of agency in analysing the ethical dimension of computational artefacts.[14] Following this definition of agency, freedom is a necessary characteristic for being an agent. Freedom allows for "intentions to act", namely internal states that determine the will to act. Johnson then argues that an artefact lacking freedom cannot be considered accountable because actions that are not freely taken are akin to reflexes or mere reactions. She also stresses, however, that the absence of freedom does not in principle preclude the inclusion of computational artefacts in the realm of morality.

Johnson suggests that computational artefacts have intentionality and that such intentionality is inherited from the artefacts' designers. Computational artefacts receive inputs and produce outputs following the instructions codified by their designers. The operations performed by these artefacts and their behaviours, are 'representations', by-products of their designers' intentions. In this context, Johnson affirms that "[computational] artefacts are prosthetic" and cannot be truly ethically autonomous as they are doomed to be always faithful images of their designers' intentions in their doing.

Johnson's argument builds on a rich definition of agency in which freedom, mental states and intention to act play necessary roles. Challenging this definition of agent allows for a different account of the morality of artificial agents. Along this line, L. Floridi and J. Sanders redefine agency in terms of input, output and transition rules.[15]

In this minimalist [16] account of agency, autonomy becomes the capability of an entity to change the rules that determine its own state transitions.[2]

Floridi and Sanders underline the fact that such autonomy depends on the Level of Abstraction (LoA) [17] at which computational artefacts are observed. In this way, the authors, as Johnson in her article, take into account the role of the designer and his knowledge of the code[3] by which the autonomy is implemented. Floridi and Sanders discriminate between the LoA at which the designer knows the algorithm by which the artefact will change its transition rules and the LoA at which the user does not know about said algorithm. At the users' LoA, the artefact is properly autonomous as it controls how it adapts to its environment.

The arguments by Johnson, Floridi and Sanders are useful to clarify a fundamental trait of the moral sphere of the computational artefacts. Despite differences in the authors' arguments and definitions, the degree of autonomy emerges as a property of computational artefacts determinant for their ethical consequences. In the next section, these consequences are further investigated.

## 3. Outsourcing of Human Capabilities to Computational Artefacts

Some computational artefacts do not need to be directly controlled by their users or designers in order to perform their activities. These artefacts are capable of initiating and controlling their (prearranged) operations without a direct intervention and therefore can be considered, in this respect, autonomous.

Examples of this type of artefact are increasingly common. In the field of robotics, for example, Spirit and Opportunity are two explorer rovers operating on Mars' surface since January 2005. Both rovers deploy an autonomous driving system capable of mapping their surroundings so they can identify obstacles to avoid. Spirit can be told where to go, but it is its navigation system that calculates how to achieve this. Other examples of systems that do not need direct intervention for producing their output include control systems, capable, for example, of flying airplanes; server applications that operate depending on the network traffic; or so called multi-agent systems used, for example, to reconfigure logistic scenarios depending on the overall system conditions.

The type of autonomy just described is weaker than the one proposed by Floridi and Sanders as self-controlled changes in transition rules are not assumed. It is simply required that computational artefacts are able to control their own interactions with the environment (namely their own operations) without direct external intervention.[4] Nonetheless, computational artefacts endowed with such autonomy have deep consequences on the environment in which they are deployed.

Autonomous artefacts allow users to outsource part of their activities. Such outsourcing represents a change in how users and computational artefacts interact that

---

[2] Machine learning algorithms are an example of this minimal autonomy.

[3] 'Code' denotes a set of instructions expressed in one or more formal language and that compose a computer program.

[4] Note that this weaker form of autonomy would not be suitable for defining agency. For a discussion on how a weaker definition of autonomy could be used for distinguishing agents from actors see [18]. For a discussion on a minimalist definition of the moral sphere of computational artifacts based on this definition of autonomy see [19].

radically departs from the purely instrumental conception seen in Wiener's review. Autonomous computational artefacts do not simply extend the capabilities of their users; they literally take over part of their activities. This way of deploying computational artefacts produces a new social dimension in which computational systems and human agents interact in parallel for achieving a given set of goals.[5]

This human/machine 'collaboration' has been expanded and globalised by the diffusion of interconnected computational systems with different degree of autonomy. The distribution at geographical level of the computational and data resources has created social, academic and commercial phenomena like the Internet and (inter)national grids. These networks have further propelled the diffusion of PCs and have also offered the infrastructures for creating a parallel social sphere – the so-called infosphere.[24]

This informational sphere is populated by inforgs [25] the informational incarnation of individuals, organisations and computational artefacts worldwide. The rising number of activities performed in the infosphere is promoting an information enrichment of the entire biosphere. Biosphere and infosphere begin to overlap as many activities, previously exclusively performed in the biosphere, are becoming informational. This phenomenon transforms traditional ethical problems and for example, as with informational privacy, crime becomes cyber-crime or economy cyber-economy.

From a technological point of view, this overlap is further propelled by several factors such as the miniaturisation, the specialisation and the increase in computational power and autonomy of networked artefacts. Altogether, these factors are shaping the next evolutionary step often indicated as 'Ubiquitous computing' or 'Pervasive computing'.[26,27] With such terms a change in the amount and type of usage of computational resources is indicated but also a novel way for society to access computational artefacts. Computation is becoming more 'transparent' to its users. Embedded into their everyday life it passes unnoticed as it does not require comprehension, direction or explicit guidance by its users. Computational resources assist at the supermarket, while paying for goods, chatting on a mobile or hunting for your favourite type of pasta.

No direct acknowledgment of the computational nature of such operations is required. What is revolutionary in pervasive computing is that an increasing number of social activities change their nature as they become mediated by computation without changing how they appear to the individuals performing them. How many telephone users know whether they are using an analogical or a digital switchboard? The way they use a telephone to make a call has not changed significantly yet digital switchboards represent a radical change in how users' data are managed, mined, duplicated and stored.

This change goes beyond the informational enrichment argued by Moor. Pervasive computational artefacts not only informationally enrich the context in which they are deployed but remain undisclosed to their users or patients. Behavioural pattern analysis is a typical example of such a trend.

Chicago's administrative council has developed one of the most extensive city surveillance systems in the USA with approximately 3000 cameras constantly recording the city's activity. This system was originally intended to assist with coordinating the response to an emergency by different crisis units. IBM is now using

[5] For an analysis of this dimension see also [20-23].

this system as a platform for developing software capable of spotting, autonomously, 'anomalous' behaviour.[28] The informational enrichment of emergency procedures has transformed the nature of the monitoring system making it a general-purpose information provider.[6] Such a system works in parallel with the police and security departments almost invisible to its patients – the citizens – but also to its users. Assuming that the system is reliable, the police unit alerted for a potential bomber at the Sears tower would initiate the security procedures independently from the source of the alarm.

This shift in the perception of computational artefacts magnifies their ethical impact. A hidden and pervasive activity cannot be critically controlled. While the users of the TPS system at the Panamanian hospital could (and probably should) have checked the calculation made by the system, it would not be practical or even possible for users or patients to check every possible implication of the operations performed for or on them by computational artefacts. In fact, users and patients might be ignorant of the involvement of computational artefacts at all or might not have the necessary expertise to assess computational behaviour. Moreover, it might not be possible to inspect an operation while it is performed by a computational artefact.

The impossibility of relying on the control of the users of computational artefacts increases the responsibility (and accountability) of those who design and develop such devices. In the following section of the paper different ways of dealing with this responsibility are reviewed.

## 4. Extending the Ethical Responsibilities of Software Designers

Traditionally, the accountability of software engineers for the safety and reliability of their designs was a problem only when developing critical control systems for sensitive industrial, military, economical or medical environments. Today, the described increase of ubiquitous computing with a growing degree of autonomy is making the problem of software reliability and safety more stringent. Moreover, as argued above, new categories of ethical concern arise as an effect of deploying these kinds of artefacts in a variety of environments. The problem is therefore how software engineers ought to face these ethical challenges.

This problem can be decomposed into three questions:

1. How can software engineers define the ethical impact of the artefacts that they will design?
2. Once the ethical implications of the artefacts' design are understood, how can software engineers shape this design so as to avoid such problems? Finally,
3. How can designers verify that the functionalities of the developed artefacts satisfy the given ethical conditions?

These three questions are related to the research area of Professional Ethics. Professional Ethics is a branch of applied ethics that investigates issues concerning the conduct of professionals working in a specific discipline. As it relates to the practice of software engineering, much of the literature focuses on the debate about the necessity and content of codes of conduct. These codes, produced by major organisations related

[6] P. Brey borrows the term "Function creep" for describing how a technology created for a specific purpose can be used for unanticipated new goals.[29] Function creep is indicated as one of the ethically problematic aspects of video surveillance technologies.

to the profession of software engineering, define a set of principles, often expressed in the form of imperatives, which practitioners should follow in their professional conduct.

In [30] it is remarked how several authors have pointed out both the practical and theoretical limits of such an approach. N. Fairweather, for example, laments the excessive focus on the traditional areas of computer ethics of privacy, accuracy, property and accessibility.[31] This restricted set of problems does not take into account relevant issues that software engineers face today such as environmental compatibility, weapon development and telecommuting.

J. Ladd also criticises the theoretical confusion between ethics and the law-like directives of the professional codes and the confusion between ethical deliberation and following directives.[32] Ladd argues that professional codes are more like a set of rules that should be followed to avoid legal problems or detrimental consequences for the professionals' career. Such a set of rules would not constitute an ethical code, as ethics should be concerned with deliberation in order to understand what choice should be made and not with acceptance of a set of directives.

Responding to these criticisms by extending the professional codes to include a wider set of ethical issues misses the point. The same applies when it is suggested that software engineers should use professional codes for critically assessing their own ethical standings. The number of ethical issues involved in designing computational artefacts is potentially innumerable and highly dependant on the specific environment in which such artefacts will be deployed. Even assuming that it might be possible to formulate an exhaustive list of ethical issues, the problem of deciding which one should be considered relevant would remain open.[7]

A true assessment of an individual's ethical standing is done by means of ethical reasoning that is not reducible to a code of conduct. Ethics is a discipline in its own, with its own conceptual 'toolkits' and specialisations. It seems unlikely that software engineers might easily or should at all be transformed into ethicists. However, it is reasonable to envision a joint team of engineers and ethicists working towards a common goal.

The ethicists of an interdisciplinary team would be best able to offer a solution for the problem of assessing the nature of the ethical impact produced by the artefact to be designed. On the other hand, the engineers would still have to solve the problem of understanding how a design might be formalised and verified taking into account such knowledge

The Values Sensitive Design (VSD) approach attempts to offer a theoretical solution for this problem. VSD is "a theoretically grounded approach to the design of technology that accounts for human values […] throughout the design process".[33] The iterative methodology proposed by VSD is based on three different investigations. The first is conceptual and aims to define the set of values relevant for a specific project. The second investigation is empirical and aims to assess the localised characteristics of the environment in which the artefact will be deployed. Finally, the third investigation is called 'technical' and it should (i) uncover how different designs support values and (ii) proactively produce a design that specifically implements the values singled out by the conceptual investigation.

[7] Note that a taxonomy or hierarchy of values would not address such a problem either as the issue here would be to understand which portion of the taxonomy is relevant for a given project.

A limitation of VSD is its lack of a precisely definable, general-purpose method for translating the knowledge accumulated in the conceptual and empirical investigations into the phase of developing the artefact itself. Such a method would be useful for avoiding 'ad hoc' solutions that would have to be implemented from scratch for every new project.

In [18, 34] a method has been proposed for translating an arbitrary number of ethical principles into preconditions for the execution of the artefact's operations. This method is compatible with VSD but overcomes the described limitation. The first step of the proposed translation method is adopting descriptive qualitative techniques for eliciting the ethical requirements relevant in the environment in which the computational artefact will be deployed.[8] In the second step, these ethical requirements are translated into the design specification by means of a conceptual tool that has been called 'control closure'.[9] Such a tool can be readily formalised by means of set theory so as to be compatible with formal approaches to design specification.

The degrees of autonomy (as defined above) and those of control are directly proportional. The more control a computational artefact has on an operation, the less external intervention is required to perform it and consequently the more autonomous the artefact is. The control closure allows for reasoning about what degree of control every component of a distributed system exercises over a given operation. For any operation performed by the system, the control closure includes the state variables necessary for its execution and also identifies to which component of the system such variables belong. Thanks to this tool, it is possible to specify the degree of autonomy that will be given to the computational artefact at a design time.

As illustrated above, the degree of autonomy is a critical characteristic in assessing the ethical impact of computational artefacts. The control closure is therefore a valuable instrument for regulating such impact at design time. In [18,34] it has been shown how given a set of ethical principles defined informally in terms of control – for example Information Privacy or safety – it is possible to use the control closure for translating these principles into a formal specification for software design.

## 5. Conclusion

This article has analysed several aspects of the relationship between the design of computational artefacts and their ethical implications.

The limits of a purely instrumental approach to computational artefacts have been clarified by pointing out the necessity of considering how users, patients and designers relate to artefacts in a social context. In this context, a role of Computer Ethics has been argued, not only as it concerns policy making for the deployment of artefacts, but also the direct participation in their design.

The analysis of how the ethical concepts of information, autonomy, control and agency are essential for understanding the nature of the ethical dimension of computational artefacts has uncovered the importance of the concept of autonomy. The ethical implications of the degree of autonomy of computational artefacts have been

[8] The methods used are mainly field-work observations with semi-structured interviews and focus group. For a description of the field-work conducted and its analysis see [35].

[9] For an exhaustive analysis of the control closure see [18, 34, 36] Please note that in [36] the control closure has been called 'ambit' for reason of technical opportunity.

assessed and it has been argued that such implications imply an increase in the responsibility and accountability of software designers. A critical analysis of how to deal with such accountability has highlighted the necessity of a truly multidisciplinary approach to software development.

The importance of devising conceptual and formal tools in order to support such multidisciplinary design teams has been stressed and the control closure has been briefly presented as a step in this direction. Further research [19,35] is currently assessing how descriptive methodologies for requirement elicitation can be adopted to uncover the ethical principles that are relevant for the deployment of computational artefacts in a given environment. Simultaneously, exploration is ongoing into the possibility of using tools as the control closure for designing computational artefacts that help to endorse ethical policies in heterogeneous organisations.

## References

[1] Bynum, T.W., *The foundation of computer ethics.* SIGCAS Computers and Society, 2000. 30(2): p. 6-13.
[2] Bynum, T.W., *Flourishing Ethics.* Ethics and Information Technology, 2006. 8(4): p. 157-173.
[3] Wiener, N., *Cybernetics: or control and communication in the animal and the machine*. 1948, New York: Wiley.
[4] Wiener, N., *The human use of human beings: cybernetics and society*. 1989, London: Free Association.
[5] Bynum, T.W., *Norbert Wiener's Vision: The Impact of "the Automatic Age" on Our Moral Lives*, *The Impact of the Internet on Our Moral Lives*, R.J. Cavalier, Editor. 2005, State University of New York Press: Albany. p. 11-25.
[6] Pugh, E.W., *Building IBM: shaping an industry and its technology*. History of computing. 1995, Cambridge, MA: MIT Press.
[7] Black, E., *IBM and the Holocaust: the strategic alliance between Nazi Germany and America's most powerful corporation*. 2002, London: Time Warner Paperbacks.
[8] Moor, J.H., *Reason, Relativity, and Responsibility in Computer Ethics.* Computers and Society, 1998. 28(1): p. 14-21.
[9] Moor, J.H., *Towards a theory of privacy in the information age.* SIGCAS Comput. Soc., 1997. 27(3): p. 27-32.
[10] Warren, S. and L.D. Brandeis, *The Right to Privacy.* Harvard Law Review, 1890. 4(5): p. 193-220.
[11] Javelin, S.a.R., *2007 Identity Fraud Survey Report*. 2007.
[12] Jakobsson, M. and S. Myers, *Phishing and counter-measures: understanding the increasing problem of electronic identity theft*. 2007, Hoboken, N.J.: John Wiley & Sons.
[13] IAEA, *Investigation of an Accidental Exposure of Radiotherapy Patients in Panama*. 2001.
[14] Johnson, D.G., *Computer systems: Moral entities but not moral agents.* Ethics and Information Technology, 2006. 8(4): p. 195-204.
[15] Floridi, L. and J.W. Sanders, *On the Morality of Artificial Agents.* 2004. 14(3): p. 349-379.
[16] Greco, G.M., et al., *How to do philosophy informationally.* Lecture notes in computer science, 2005. 3782: p. 623-634.
[17] Floridi, L., *The Method of Abstraction*, in *Yearbook of the Artificial dedicated to "Models in contemporary sciences"*, P. Lang, Editor. 2004. p. 177-220.
[18] Turilli, M., *Ethical Protocols Design.* Ethics and Information Technology, 2007. 9(1): p. 49-62.
[19] Turilli, M., *Ethical outsourcing in computational artefacts.* Forthcoming.
[20] Latour, B., *Reassembling the Social: An Introduction to Actor-Network-Theory*. 2005, Oxford: Oxford University Press.
[21] Verbeek, P. and P. Kockelkoren, *The Things That Matter.* Design Issues, 1998. 14(3): p. 28-42.
[22] Winner, L., *Do Artefacts Have Politics?* 1986: University Of Chicago Press.
[23] Verbeek, P., *What Things Do: Philosophical Reflections on Technology, Agency, And Design*. 2005: Pennsylvania State University Press.
[24] Floridi, L., *On the Intrinsic Value of Information Objects and the Infosphere.* Ethics and Information Technology, 2002. 4(4): p. 287--304.
[25] Floridi, L., *A Look into the Future Impact of ICT on Our Lives.* The Information Society, 2007. 23(1): p. 59-64.

[26] Satyanarayanan, M., *Pervasive computing: vision and challenges.* IEEE Personal Communications, 2001. 8(4): p. 10-17.
[27] Weiser, M., *Some computer science issues in ubiquitous computing*. 1999, ACM Press. p. 12.
[28] IBM, *The City of Chicago's OEMC and IBM Launch Advanced Video Surveillance System*, in *IBM Press Releases*. 2007.
[29] Brey, P., *Ethical Aspects of Facial Recognition Systems in Public Places*, in *Readings in cyberethics*, R.A. Spinello and H.T. Tavani, Editors. 2001, Jones and Bartlett Publishers: Boston. p. 585-600.
[30] Spinello, R.A. and H.T. Tavani, *Readings in cyberethics*. 2001, Boston: Jones and Bartlett Publishers.
[31] Fairweather, N.B., *No, PAPA: Why Incomplete Codes of Ethics are Worse than None at All*, in *Ethics in an Age of Information Technology*, G. Collste, Editor. 1998, New Academic: Delhi.
[32] Ladd, J., *The quest for a code of professional ethics: an intellectual and moral confusion*, in *Ethical issues in the use of computers*. 1985, Wadsworth Publ. Co. p. 8-13.
[33] Friedman, B., P. Kahn, and A. Borning, *Value Sensitive Design: Theory and Methods*. 2002, University of Washington: Technical report.
[34] Turilli, M., *Ethical flexibility for artificial agents.* Forthcoming.
[35] Turilli, M., *Fieldwork for Ethical Requirement Elicitation in the VOTES project.* Forthcoming.
[36] Sanders, J.W. and M. Turilli. *Dynamics of Control*. in *TASE 2007*. 2007.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*


# How to Explain the Underrepresentation of Women in Computer Science Studies

Margit POHL[a], Monika LANZENBERGER[b]
[a]*Institute of Design and Assessment of Technology, Vienna University of Technology*
[b]*Institute of Software Technology and Engineering, Vienna University of Technology*

**Abstract**. An overview of recent research concerning the underrepresentation of women in computer science studies indicates that this problem might be more complex than previously assumed. The percentage of female computer science students varies from country to country, and there is also some indication that gender stereotypes are defined differently in different cultures. Gender stereotypes concerning technology are deeply embedded in the dominant culture and often contradictory. Only a few general assertions can be made about the development of the inclusion or exclusion of women from computer science. In addition, there does not seem to be a specific female style of computer usage. Concepts of diversity and ambivalence seem to be more appropriate but difficult to realize. All this makes the development of appropriate interventions for overcoming the underrepresentation of women in computer science studies a very complex process.

**Keywords.** Gender, computer science, university education, diversity

## Introduction

The underrepresentation of women in computer science studies has been investigated extensively in the past decades. The results of these investigations are somewhat contradictory, especially if we consider not only the development in industrialized Western countries but also in Third World countries. In some industrialized countries there was an increase in the percentage of female computer science students in the 1970s and early 1980s and a decrease in the late 1980s and 1990s (see e.g. Austria: [1]; Germany: [2]; Great Britain: [3]; Norway: [4]; Sweden: [5]; USA: [6]). In some countries, there was a slight increase again around and after the year 2000 [7,8]. It should be mentioned, however, that this is not a global phenomenon but restricted to several countries. Wright [9] points out that this development cannot be observed in all countries. She investigated women's percentages of graduates in mathematics and computer science and distinguished between three different groups of countries:
1. Percentages rising then falling (e.g. USA, Canada, Austria)
2. Percentages falling then falling (e.g. Italy, Finland, Portugal, Hungary)

3. Percentages rising then rising (e.g. Jordan, Greece, Poland, Malaysia, Bulgaria)
This indicates that the explanation for the (lack of) participation of female students in computer science studies is rather difficult, especially if we look at it globally. This will be discussed in more detail below.

Numerous studies have tried to explain the relationship between gender and computer technology. There seems to be a broad consensus that female and male attitudes toward the computer have to be seen in conjunction, and that one cannot be explained without considering the other. Gürer and Camp [10] developed a very comprehensive framework of reasons for the underrepresentation of women in computer science, among them women's negative attitude towards computers, the fact that computer games are predominantly developed for boys, and the lack of equal access for girls to computers. Schinzel [11] discussed a few additional explanations, especially that informatics shifted from a discrete application of mathematics to an engineering discipline. She argues that this made it more difficult for women to identify with computer science. Grundy [12] also supports this view. Erb [13] found out in her interviews that female computer scientists feel especially attracted by the mathematical side of computer science. This argument is possibly specific for Germany. In contrast to that, Margolis and Fisher [14] assume that the similarity of computer science to mathematics makes it especially difficult for women to get interested in this subject, a view which is also supported by Grundy [12].

Margolis and Fisher [14] posit that an important reason for the underrepresentation of women in computer science is the "geek mythology" prevailing among computer scientists. This view is also supported by Rasmussen [15] who argues that young women are deterred from computer science by the image of the compulsive programmer or computer nerd. They think that computer scientists are "someone else – different from us" ([15], p.385). Margolis and Fisher also mention lack of access, computer games and lack of confidence as important reasons for the low percentage of women in computer science studies. They describe interventions to increase the number of female students in computer science, among them change of admission policy, change of curriculum and contextualizing computer science by placing computers in the context of their real-world uses. It is well known that their policy was rather successful, but some of their interventions have been criticized recently [16]. Blum et al argue that it would be too narrow a view just to increase the number of women in computer science because this would suggest that only women have an ambivalent relationship to computer science "as it is". By changing general admission criteria not only the percentage of women can be increased, but also different groups of men are addressed - men who do not conform to the "geek mythology", or who belong to specific ethnic groups. Less emphasis on prior programming experience and more importance for leadership promise are examples of such modified admission criteria at Carnegie Mellon.

The discussion about the experiences at Carnegie-Mellon University indicates that the controversy about possible explanations for the underrepresentation of women in computer science is important because the nature of interventions chosen for overcoming this problem depends on how the underrepresentation is explained. If the explanation for the underrepresentation of women in computer science is seen as lying in the "nature" of women (e.g. women prefer more practical and application oriented areas of computer science and are deterred by mathematics), then curricular changes adapted to women's needs seem to be an appropriate measure. As described above, this view is quite controversial, and it is an open question whether women really detest

more theoretical areas of computer science. If, on the other hand, the underrepresentation of women is seen as part of a bigger problem of the exclusion of various groups from computer science (e.g. exclusion of some ethnic groups or of people with a low level of education), then a more general approach has to be taken as Blum et al suggest.

In addition, we would like to point out that several of the reasons given for the underrepresentations of women in computer science studies (see e.g. [10] do not hold anymore. Several studies indicate that lack of access to computers is not such a problem for women anymore (see e.g. the results of our own study below and [17]). It has also been argued that computer games designed for boys/men exclude women. In contrast to this, women have increasingly used computer games in recent years, but there remains an observable gender gap (see e.g. [18]). Apart from that, a computer game like Barbie Fashion designer which specifically aims to fulfill young girls' desires was extremely successful (see e.g. [19]). The success of this game which reinforces gender stereotypes also does not fit the argumentation that girls are excluded from computer culture because of the "male" character of computer games. It is also controversial whether women really have negative attitudes towards computers. Wagner points out that those women who do work in the field of computer science often feel highly attracted to computers ([20], see also [21]). It seems that all these phenomena do not really explain the underrepresentation of women in computer science. It might be argued that these were not the reasons for the underrepresentation of women in computer science and that other reasons have to be found. We would rather assume that discrimination of women is a more complex phenomenon which has to be viewed in a holistic manner and which is, as Blum et al [16] argue, embedded in culture.

It can be concluded that there is no simple answer to the question of the underrepresentation of women in computer science studies. Some of the reasons given for this do not hold anymore. Other explanations (like, e.g., assuming that women are sceptical about mathematics) are controversial. This makes decisions about the introduction of interventions to increase the number of female computer science students very difficult. It seems that single measures are not a very promising strategy. On the other hand, it does not seem to be realistic to use a highly comprehensive approach because this would exceed the capacity of most bodies concerned with such issues. In addition, the character of the interventions also depends on the theoretical assumptions taken by the stakeholders involved (e.g. explaining the underrepresentation as something related to the "nature" of women or as embedded in the cultural environment). Theoretical considerations, whether done explicitly or implicitly, strongly influence the character of the measures for inclusion which are adopted. Therefore, it seems to be sensible to take such considerations into account when discussing interventions for increasing the number of women in computer science studies.

The following paper will first present some results of two studies we conducted at the Vienna University of Technology concerning these questions. Then we discuss the topic of the cultural embeddedness of gender differences and the controversial question whether there is something like a "female" style of doing computer science. Based on that, we will revisit the issue of formulating successful strategies for increasing the number of female students of computer science.

**Table 1.** "Do you think mathematical ability is a precondition for your studies?"

| | High | Moderate | None | No Answer |
|---|---|---|---|---|
| Female | 70.73 | 21.95 | 2.44 | 4.88 |
| Male | 65.24 | 32.62 | 1.29 | 0.86 |

## 1. Female Students of Computer Science at Vienna University of Technology

At the Vienna University of Technology, we conducted two studies to investigate the motivation and approaches of female and male students of computer science. The first study took place in 1993 with 42 female and 204 male students (=246 students). The second study took place in 2004 with 41 female and 247 male students (=288 students). The studies were based on two questionnaires (one for 1993, one for 2004) which were very similar. This allowed us to look for changes over time between 1993 an 2004. The following text describes the results of the 2004 study, and only if necessary refers to the study of 1993.

An important part of the questionnaire was concerned with abilities necessary as a precondition for studying computer science. Both female and male participants have similar opinions about necessary abilities of computer science students. In particular, we asked them about the need for technical ability and the need for mathematical ability as prerequisites for their specific studies.

46 percent of the female and nearly 42 percent of male students said a high amount of technical ability is necessary. A similar number of people, 39 percent of the female and nearly 50 percent of the male students, meant that students would need just a moderate amount of technical abilities. Almost no one (about 2 percent females and 1 percent males) thought that technical ability is not required.

When asked about mathematical ability the answers are different (see table 1). The participants judged mathematical abilities significantly more important than technical abilities. This contradicts Grundy's [12] and Schinzel's [11] assumption that computer science is seen as gradually becoming an engineering science.

In our study we also asked for the individual interest in mathematics in school (see table 2). A surprisingly large number of female students said they were highly interested in mathematics in school. Most of the male students said their interest in mathematics was moderate.

The importance of mathematical ability in combination with the high interest in mathematics seems to be a door opener for many female students. This is significantly different to programming which mainly seems to attract male students. Female and male students judge their programming skills rather differently (see table 3).

**Table 2.** “Rate your individual interest in mathematics in school"

| | Highly | Moderately | Not At All |
|---|---|---|---|
| Female | 48.78 | 34.15 | 17.07 |
| Male | 37.8 | 46.75 | 15.45 |

**Table 3**: "Do you think that you are good in programming?"

| | Yes | Moderately | No | No Answer |
|---|---|---|---|---|
| Female | 9.75 | 24.39 | 53.65 | 12.19 |
| Male | 36.03 | 44.53 | 17.0 | 2.42 |

Male students of the first semester mostly seem to be untroubled by concerns about their programming skills, which is a significant difference with the female students. However, without additional information it is difficult to predict consequences of this difference. Therefore, it is interesting to know that both genders think programming is important in their individual studies (see table 4). Obviously, this estimation affects female students in another way than their male colleagues. Lacking an important skill at least implies higher pressure and reduced self-confidence for the female students. Since computer science still has a strong programming connotation, the increased attraction of male students is evident. Following the study results of Margulies and Fisher [14] at the Carnegie Mellon University, the programming skills of freshman and freshwomen are not relevant because they do not represent potentially successful students. This is contradictory to the image of computer science studies which many students perceive as programming adventures for geeks.

Our study in 2004 showed that nearly 40 percent of the female students decided to study computer science shortly before enrollment, and another 34 percent made this decision within the last year. This differs from the behavior of the males: around 25 percent chose computer science some longer time ago. However, most of the male students (around 34 percent) decided for their specific studies shortly before enrollment as well.

Once they decided for computer science, we can see a significant difference between female and male students as to whether or not they began their studies together with friends or colleagues. In 2004 only 19 percent of the female but 44 percent of the male students registered together with friends or colleagues. It seems that males tend to choose their bachelor studies in groups with others, whereas females mostly start their favored computer science studies alone. This relation between genders has not changed within the last 11 years, although there is an increase in the tendency to begin to study in groups in both genders in recent years. The same question asked in 1993 showed that 8 percent of the females and 33 percent of the males began their studies together with friends or colleagues.

We found major changes within the last decade as well: For computer science students owning a computer seems to be a must before they begin their studies today. This holds for female students even more than for male students. In 2004 the outstanding number of 100 percent of the female and 97 percent of the male students declared that they had owned a computer before they began to study their computer

**Table 4:** "Do you think that programming is very important in your studies?"

| | Yes | No | No Answer |
|---|---|---|---|
| Female | 85.37 | 12.2 | 2.44 |
| Male | 74.18 | 16.72 | 8.2 |

science bachelor. The study in 1993 showed another picture. 83 percent of the male students owned a computer prior to their enrollment in contrast to only 48 percent of the females. In contrast to the early 1990s, access does not seem to be a problem anymore for women. And both genders use their computers often during leisure. In 2004 nearly all of the females (95 percent) and males (98 percent) said that they spend a lot of their spare time with the computer.

This corresponds to students' opinion about the relevance of private computer usage which increased considerably within the last decade. Whereas in 1993 slightly more than two-thirds (68% female and 73 % males) stated that using the computer for private purposes as well is a necessary precondition for studying computer science, this number rose to 95 percent of the females and 89 percent of the males in 2004.

This study indicates that some of the reasons given for the underrepresentation of women in computer science either do not hold anymore or depend on culture. Lack of access to computers is not a problem anymore, and, in contrast to some authors, mathematics does not deter women in Austria from studying computer science. In our study, we could substantiate other hypotheses, as e.g., that female students possess less self-confidence concerning programming than male students. The social environment seems to play an important role in women's decisions to study computer science. We think that, in general, the topic is quite complex and governed by historical change.

## 2. Discussion

### *2.1. Cultural differences*

As mentioned above, one of the reasons that no simple explanation for the underrepresentation of women in computer science studies is possible is that there are distinct regional differences, although in most countries the percentage of women studying computer science is below 50%. As Schinzel [8] remarks, it is difficult to compare the results of statistical material from different countries, but nevertheless, some conclusions can be drawn. It may be surprising that the percentages of women studying computer science or mathematics in Third World countries (as e.g. Kuwait, Brazil, Mexico) is often much higher than in industrialized countries. In Europe, the percentage of university degrees in mathematics and computer science awarded to women is very low in countries like Austria, the Czech Republic, Hungary, the Netherlands, the Slovak Republic and Switzerland. It is much higher in Mediterranean countries (Portugal, Spain, France) or Scandinavian countries (Sweden, Finland) [8]. Schinzel argues that industrialization took place much later in the Mediterranean countries, therefore the social status of engineers is lower in these countries, and it is more acceptable for women to study such subjects. These results are supported by statistics published by the European Union [22], which show that the number of female graduates in science, mathematics and computing are very high in Mediterranean countries and lower in countries like Denmark, Germany, the Netherlands and Austria. In Arabian countries, religion seems to be the most prestigious subject for study therefore it is easier for women to enter engineering, which is not as prestigious.

There is some indication that Asian women have a different attitude concerning computer science [23,24]. Lagesen describes the situation of female computer scientists in Malaysia. She argues that computer science is not seen as a predominantly masculine domain because this work takes place indoors. An office is seen as an

acceptable and protected place for a woman. Lagesen concludes that the co-construction of gender and computer science might be a more complex process than previously assumed. Greenhill et al [24] also investigated the attitudes of female students of computer science of Asian origin. In Australia, about half of the female students have an Asian background. Female students of computer science with such a background seem to have a notably different attitude towards computers. For them, the likelihood of getting a good job within computer science is a much stronger motivating factor than for other female students in this area. They are less likely to study computer science because of a deep interest in this discipline.

It seems that the concepts of femininity and masculinity vary depending on the cultural context in which they are embedded. They are fluid, easily adaptable to the environment they are set in. The only constant feature is that masculinity is always seen as something superior. If technology and engineering convey a high status, women are usually excluded from these areas. In other cultures, women are accepted into computer science more easily. The case of Malaysia as described by Lagesen is probably not a best practice scenario that can be adapted for other countries, and it seems that no universal strategies for improving the situation of women in computer science can be developed. Moreover, it can be concluded that the low representation of women in computer science must always be seen as a phenomenon related to a specific cultural environment. The cultural embeddedness of discrimination against women in technology is probably also the reason for the lack of results of interventions in this area which is sometimes reported. Despite a policy supporting the inclusion of women into technology, there is fairly little progress made, as the case of computer science studies in many industrialized countries demonstrates. It might be argued that this policy can only foster a fast and thorough process of change when the attitudes of all the major stakeholders in this process and more generally cultural habits in the society are changed.

### *2.2. "Female style" vs. diversity*

Another theoretical issue influencing the nature of interventions to overcome the underrepresentation of women in computer science studies is the question whether there is some specific female style in computing. If we assume, for example, that there is a specific female programming style which Turkle (see e.g. [25]) would call soft mastery und which is characterized by an intuitive and holistic approach, then it would be plausible to develop a special curriculum for women. Schelhowe [26] argues that there is empirical evidence contradicting Turkle's assumption of a specific female programming style, and, based on such evidence, Blum et al [16] reject the idea of a curriculum based on women's needs.

In general, it is extremely difficult to define what a female and a male programming style or style of computer usage might be. Faulkner [27] points out that there are at least two different "masculine" approaches, one which is defined as an abstract, top-down strategy and another one (the "hacker" approach) which is more bottom-up, including trial and error. The second approach is very similar to what Turkle defines as soft mastery. Everyday experience leads us to believe that gender identities are fairly stable and well-defined. If analyzed in detail, however, they turn out to be highly flexible and contradictory.

To avoid the pitfalls of an essentialist approach, the concepts of diversity (see e.g. [28]) and ambivalence [29] were developed. The concept of diversity is based on the

notion that there are many different approaches to computer usage based on gender, class, ethnicity, age and other variables. Computer software has to be designed in a way to support all these different approaches. It has to be open, flexible and adaptable, so that every user can accommodate his or her own style ("undetermined design"). This approach seems to be very interesting, but it might also be dangerous. Wagner [20] argues that the concept of diversity is very tempting, but the analysis of the discrimination of women still affords the definition of "women" as a universal category. Otherwise the discrimination of women might disappear in a multitude of different views and approaches. Gender as a topic of scientific research would, in the long term, disappear.

Ambivalence [29] is a similar concept. Collmer assumes that there is no specific female style of computer usage, but that their approach to computers is ambivalent. On the one hand, women are influenced by the culture of exclusion of women from technology. On the other hand, women are often interested in technology, and also, if working in that area, forced to adapt to its (social) rules and regulations. This causes feelings of ambivalence in them. Collmer could show empirically that feelings of ambivalence were much less pronounced in men than in women.

To conclude, there seems to be little empirical evidence for the assumption that there is a specific female style of computer usage. The alternative to such an approach are concepts like diversity or ambivalence which are more complex and promising.

## 3. Conclusion

An overview of recent research concerning the underrepresentation of women in computer science studies indicates that this problem might be more complex than previously assumed. Some of the explanations given for this phenomenon apparently are not valid anymore. The fact that the extent of the underrepresentation of women in computer science studies is dependent on cultural factors and that there is apparently no specific female style of computer usage make it difficult to identify appropriate measures to overcome this problem. Best practice models from other countries might not be applicable because of cultural differences. The diversity approach seems to be very promising but difficult to achieve. Although there is a long tradition of research in the area of underrepresentation of women in computer science studies, more research in this area seems necessary. We think that refining the concepts of diversity and ambiguity and discussing their implications for practical measures seems to be a very interesting area of research in the near future.

## Acknowledgements

We want to thank Margit Schütz, Emine Kara, Selva Ardic and Markus Rester for their help in the empirical investigation of 2004.

## References

[1] Pohl, M. (1997) Beruf oder Berufung: Zur Situation der Informatikerinnen an der Technischen Universität Wien. In: J. Mikoletzky, U. Georgeacopol-Winischhofer, M. Pohl: “Dem Zug der Zeit

entsprechend...” Zur Geschichte des Frauenstudiums in Österreich am Beispiel der Technischen Universität Wien. Wien: WUV Universitätsverlag, p.301-324
[2] Behnke, R., Oechtering, V. (1995) Situations and Advancement Measures in Germany. In: Communications of the ACM. January 1995/Vol.38, No.1, p.75-82
[3] Kirkup, J. (1992) The Social Construction of Computers: Hammers or Harpsichords? In: J. Kirkup, L. Smith Keller (eds.) Inventing Women. Cambridge, UK, Polity Press, p.267-281
[4] Berg, V. A. L., Gansmo, H.J., Hestflått, K., Lie, M., Nordli, H., Sørensen, K.H. (2002) Gender and ICT in Norway: An overview of Norwegian research and some relevant statistical information. Report 02/Part 4 of the IST-2000-26329 SIGIS project; http://www.rcss.ed.ac.uk/sigis/public/documents/SIGIS-D02_Part4.pdf (last seen: 23.09.2007)
[5] Björkman, C., Christoff, I., Palm, F., Vallin, A. (1997) Exploring the Pipeline: Towards an understanding of the male dominated computing culture and influence on women. In: R. Lander, A. Adam (eds.) Women in Computing. Exeter, England: intellect, p.50-59
[6] EECS Women Undergraduate Enrollment Committee (1995) Women Undergraduate Enrollment in Electrical Engineering and Computer Science at MIT. Final Report of the EECS Women Undergraduate Enrollment Committee. January 3, 1995
[7] Lanzenberger, M., Pohl, M. (2005) Media Informatics or Software Engineering: Why do Women Study Computer Science? In: 6th International Women into Computing Conference (WiC), p.266-274
[8] Schinzel, B. (2005) Kulturunterschiede beim Frauenanteil im Informatik-Studium. http://mod.iig.uni-freiburg.de/cms/index.php?id=173 (last seen: 23.09.2007)
[9] Wright, R. (1997) Women in Computing: A Cross-National Analysis. In: R. Lander, A. Adam (eds.) Women in Computing. Exeter, England: intellect, p.7283-59
[10] Gürer, D., Camp, T. (2001) Investigating the incredible shrinking pipeline for women in computer science. Final Report – NSF Project 9812016
[11] Schinzel, B. (1997) Why has female participation in German Informatics decreased? In: Women, work and computerization. Spinning a web from past to future. Proceedings of the 6th International IFIP Conference, p.365-378
[12] Grundy, F. (1998) Mathematics in Computing: A Help or Hindrance for Women? http://www.cs.keele.ac.uk/content/people/a.f.grundy/maths.htm (last seen 23.09.2007)
[13] Erb, U. (1996) Frauenperspektive auf die Informatik. Münster: Westfälische Dampfboot
[14] Margolis, J., Fisher, A. (2003) Unlocking the Clubhouse. Women in Computing. Cambridge, London: MIT Press
[15] Rasmussen, B. (1997) Girls and Computer Science: "It’s not me. I’m not interested in sitting behind a machine all day.“ In: Women, work and computerization. Spinning a web from past to future. Proceedings of the 6th International IFIP Conference, p.379-386
[16] Blum, L., Frieze, C., Hazzan, O., Dias, M.B. (2006) A Cultural Perspective on Gender Diversity in Computing. (long version of a paper presented at SIGCSE 2006) http://www.cs.cmu.edu/~lblum/PAPERS/CrossingCultures.pdf (last seen: 23.09.2007)
[17] Sørensen, K. (2002) Love, Duty and the S-Curve. An overview of some current literature on gender and ICT. Report 02/Part 1 of the IST-2000-26329 SIGIS project; http://www.rcss.ed.ac.uk/sigis/public/documents/SIGIS-D02_Part4.pdf (last seen: 23.09.2007)
[18] Hartmann, T., and Klimmt, C. (2006). Gender and computer games: Exploring females' dislikes. Journal of Computer-Mediated Communication, 11(4), article 2. http://jcmc.indiana.edu/vol11/issue4/hartmann.html
[19] Subrahmanyam, K., Greenfield, P.M. (1998) Computer Games for Girls: What makes them play? In: J. Cassell, H. Jenkins (eds.) From Barbie to Mortal Combat. Gender and Computer Games. Cambridge, Mass., London, England: MIT Press, p.46-71
[20] Wagner, I. (1992) Feministische Technikkritik und Postmoderne. In: I. Ostner, K. Lichtblau (eds.) Feministische Technikkritik. Ansätze und Traditionen. Frankfurt, New York: Campus, p.147-163
[21] Corneliussen, H. (2005) Women's Pleasure in Computing. In: 6th International Women into Computing Conference (WiC), p.251-265
[22] European Commission (2003) She Figures 2003. http://ec.europa.eu/research/science-society/pdf/she_figures_2003.pdf (last seen 23.09.2007)
[23] Lagesen, V., Mellström, U. (2004) "Why is computer science in Malaysia a gender authentic choice for women? Gender and technology in a cross-cultural perspective", In: Professional Learning in a Changing Society. International Conference, Nov.2004 http://www.pfi.uio.no/konferanse/prof_learning/docs/pdf/session3/Lagesen.pdf (last seen 23.09.2007)
[24] Greenhill, A., Hellens, L.von, Nielsen, S., Pringle, R. (1997) Australian Women in IT Education: Multiple Meanings and Multiculturalism. In: Women, work and computerization. Spinning a web from past to future. Proceedings of the 6th International IFIP Conference, p.387-397

[25] Turkle, S. (1986) Die Wunschmaschine. Der Computer als zweites Ich. (engl. "The Second Self") Reinbek bei Hamburg: Rowohlt
[26] Schelhowe, H. (1999) Interaktivität der Technologie als Herausforderung an Bildung. Zur Gender-Frage in der Informationsgesellschaft. In: Jahrbuch Arbeit, Bildung, Kultur Bd. 17, 1999, p.49-55
[27] Faulkner, W. (2000) The Power and the Pleasure? A Research Agenda for "Making Gender Stick" to Engineers. In: Science Technology Human Values 2000; 25, p.87-119
[28] Cassell, J. (2003) Genderizing Human-Computer Interaction. In: J.A. Jacko, A. Sears (eds.) The Human-Computer Interaction Handbook. Mahwah, New Jersey: Lawrence Erlbaum, p.401-412
[29] Collmer, S. (1997) Frauen und Männer am Computer. Wiesbaden: Deutscher Universitätsverlag

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*


# How the Web Is Changing the Way We Trust

Dario TARABORELLI[1]
*University College London*
*d.taraborelli@ucl.ac.uk*

**Abstract.** Several studies have addressed the issue of what makes information on the *World Wide Web* credible. Understanding how we select reliable sources of information and how we estimate their credibility has been drawing an increasing interest in the literature on the Web. In this paper I argue that the study of information search behavior can provide social and cognitive scientists with an extraordinary insight into the processes mediating knowledge acquisition by epistemic deference. I review some of the major methodological proposals to study how users judge the reliability of a source of information on the *World Wide Web* and I propose an alternative framework inspired by the idea that–as cognitively evolved organisms–we adopt strategies that are as effortless as possible. I argue in particular that Web users engaging in information search are likely to develop simple heuristics to select in a cognitively efficient way trustworthy sources of information and I discuss the consequences of this hypothesis and related research directions.

**Keywords.** Credibility, authority, trust, heuristics, information search

## 1. Judging Epistemic Reliability on the *World Wide Web*

Possessing reliable knowledge and being able to select reliable sources of information are skills essential to our capacity to cope in an efficient way with the problems raised by our physical and social environment. As evolved cognitive organisms, we negotiate demanding cognitive problems by selecting parsimonious strategies that provide us with sufficiently accurate solutions. *Epistemic deference*–or the ability to trust external sources of information to form new beliefs–can be regarded as one such strategy. In this paper I suggest that epistemic deference is a common aspect of information search on the *World Wide Web* and I argue that in order to be cognitively efficient it has to rely on simple and relatively effortless heuristic strategies.

[1] I am grateful to the participants in E-CAP 2007 for valuable feedback and discussions on an earlier version of this paper. This work was partly supported by a Marie Curie fellowship from the European Commission (MEIF-CT-2006-024460). *Correspondence address:* Department of Psychology, University College London, Gower Street, London WC1E 6BT, United Kingdom.

### *1.1. Epistemic deference and the problem of selecting reliable sources*

Social epistemology has introduced the concept of "epistemic deference" to refer to those processes of belief formation in which an agent (the *deferrer*) relies on an external source (the *deferee*) in order to extend her knowledge to facts with which she has no direct acquaintance, or more generally, to use information from this source as "a model for what to believe".[1] Relying on experts in order to make a decision is a typical example of a process in virtue of which we adopt a *deferential stance* towards another agent's opinions (an *epistemic authority*) to extend our system of beliefs beyond its individual boundaries. Epistemic deference is a constitutive trait of language competence, as the capacity by which we can "entertain thoughts through the language that would not otherwise be accessible to us" [2] as when – for example – we use the term "arthritis" in a conversation without exactly knowing the precise reference of this term. But the scope of deference is arguably broader than language use. Deference to an external source of information, insofar as we trust that source for epistemic matters, allows us to extend our beliefs and our ability to reason about facts which we do not thoroughly master.[3] As such, deference is a principle found everywhere in human cognition and possibly one of the common strategies used to bootstrap knowledge and language acquisition in young children. (e.g, [4])

In general, we defer to external sources of information: (i) whenever we lack reliable knowledge on a given subject matter to ground a decision (in which case deference is a *necessary* condition) or (ii) when deference provides a convenient, parsimonious *sufficient* solution to meet the requirements of a problem.

a) If I am not a doctor, looking up symptoms of a disease in a medical encyclopedia can be regarded as an example of the first situation: trusting medical information about a disease from an encyclopedia is a *necessary* form of epistemic deference since I am not a medical expert and I could not acquire knowledge about this disease if not by deferring to an expert source.
b) Now consider the case in which I cannot directly recall a friend's phone number and I decide to call another friend who may know this number by heart: in this case, I trust the person that I am calling (or rather, her memory) as a *sufficiently* reliable source to provide the information that I need, even if I could directly check this information by myself by other means. For example I could go back home and find the number in the copy of the address book I keep near my landline phone.

The massive availability of information on the *World Wide Web* is making deferential practices as those exemplified by (a) and (b) a constitutive part of our belief formation and decision making strategies. Searching the Web typically yields multiple sources for the piece of information we are looking for, so the critical question we face is which of these sources should be trusted.

How do we select *trustworthy deferees* when we engage in information search on the Web? Unsurprisingly, it has been shown that we systematically rely on background knowledge and previous experience as main factors to decide whether a source of information on the *World Wide Web* is trustworthy or not. Familiarity and *experienced credibility* [5] are among the most common grounds for the selection of trustworthy

sources to which we defer. In the general case, though, we have no prior information on the trustworthiness of an external source and we need to estimate it.

### *1.2. Evaluative judgments of epistemic reliability*

The problem of credibility of electronic information has been the object of a growing body of literature in the last decades.[6] Studies of Web credibility [5,7] or the perception of epistemic trustworthiness of unfamiliar sources on the Web have mostly focused on *evaluative judgments*, i.e. judgments people make in order to estimate the trustworthiness of a source of information on the basis of extensive inspection of the content and credentials provided by the source. Evaluative judgments should be distinguished by *predictive judgments* or judgments about the expected reliability of a source prior to its actual inspection. [8]

I will return later to the *evaluative vs. predictive* distinction, but it is worth asking why mainstream research on Web credibility has been focusing on evaluative judgments. Arguably, the main reason why the study of evaluative judgments has been privileged in the literature is that in ideal conditions, whenever agents are required to estimate the credibility of a source, they are not subject to particular constraints of time or cognitive effort to make this judgment.

Traditionally, the study of evaluative judgments of credibility of a source of information has addressed two central issues:

1. how easily information acquired by deferring to external sources can be integrated into one's system of beliefs;
2. how prone such information is to subsequent revision.

Mainstream theories of persuasion [9,10,11] suggest that among the factors affecting the likelihood of subsequent revision of an evaluative judgment, the *amount of processed information* and the *degree of involvement* play critical roles. Judgments based on small amounts of information or in conditions of low involvement are more likely to be subsequently revised [12,13]. It would then seem natural to assume that in the case of knowledge acquisition mediated by information search, people are likely to invest a large amount of information processing effort with the goal of identifying trustable sources of information. I will try to show that this assumption cannot be taken for granted.

Web credibility studies have indeed collected large datasets on evaluative judgments of source reliability by asking users to rate the quality of visited websites. [14] The results single out several factors affecting:

1. the likelihood that specific features of the source be noticed by the subject;
2. the attribution of positive or negative values to features that are noticed.

Verbal reports and qualitative, questionnaire-based methods have been the most popular approach to studying judgments of credibility and source reliability at least since two decades [15] and are still the dominant approach adopted to investigate credibility on the *World Wide Web*. It should be noted that this approach is not limited to judgments of the credibility of Web sources decontextualized from the specific task in which such judgments are required: even when researchers look at judgments of epistemic reliability in real human-computer interaction tasks, they tend to privilege

verbal reports (or "think-aloud" protocols) as the main source of empirical evidence over other possible kinds of behavioral data [16].

### *1.3. Beyond qualitative analyses of epistemic reliability judgments*

The use of verbal reports to understand Web credibility relies on the assumption that introspection is the best way to determine factors affecting judgments and decisions on the epistemic reliability of a source. As, for example, [16] observes,

> the method used in this study is premised on the assumption that the users can identify and discuss the characteristics and features of information objects that influence their judgments of information quality and cognitive authority. (p.150)

However useful verbal reports may prove to the study of evaluative judgments in decontextualized conditions, they face a number of major limitations:

1. verbal reports assume that subjects are aware of the factors affecting the selection of a specific source as credible or epistemically reliable, but there is no need to assume that processes involved in such judgments should be explicit;
2. reports relying on extensive inspection of a Web source can hardly account for the kind of processes in which users engage when they are involved in real-world information search tasks, which are usually constrained by time and by limits on the cognitive effort the user is willing or able to invest in the task;
3. qualitative studies based on in-depth, decontextualized source evaluation implicitly take for granted that *a posteriori*, evaluative judgments are immune to task dependence effects (which may affect a user's perceived utility of the different features and credentials of a source);
4. evaluating credibility judgments against a list of predefined dimensions on the basis of verbal reports typically skews the results in favor of an (often arbitrary) class of credibility variables chosen by the experimenter.

These methodological issues affecting qualitative studies of Web credibility call for an alternative framework to study user behavior in ecologically valid conditions and under the typical constraints of real information search tasks.

### *1.4. Predictive judgments of epistemic reliability*

Information foraging studies [17–20] marked a turning point in the literature on information search, by drawing the attention of the research community to the importance of studying *predictive judgments* of the value of a source of information in the context of an information search task as opposed to *a posteriori* evaluations. Predictive judgments are those that users make when they evaluate a source on the basis of information describing a source, such as link descriptions. This "proximal information" can allow the user to estimate which source to visit prior to its in-depth evaluation.

Why are predictive judgments underrepresented in current research on Web credibility? One of the possible reasons is that low effort judgments may have been ruled out as irrelevant to the understanding of trust and epistemic deference. Since beliefs formed in low effort condition (i.e. those processes that tend to be classed as

"peripheral routes to persuasion" [10]) have been shown to be more volatile and less predictive of behavior, researchers may have erroneously assumed that these are not as representative of deferential behavior as beliefs acquired on the basis of in-depth, more cognitively demanding evaluations.

The question boils down to understanding what the average level of engagement typical of information search behavior on the Web is. As Fogg observes:

> Web users typically spend small amounts of time at any given site or individual page, and are thus likely to develop strategies for assessing credibility quickly. One could argue that people typically process Web information in superficial ways, that using peripheral cues is the rule of Web use, not the exception. From a user perspective, there are too many competitors on the Web for deep credibility evaluation. Even the words people use to describe Web use–"visiting sites" and "surfing the Web"–suggest lightweight engagement, not deep content processing. Research has yet to examine the relationship between engagement level and credibility assessments online. [7]

It is reasonable to assume that, depending on the goals of the information search task, there may be different possible degrees of deference to a source ([1], p.189), and–consequently–different degrees of cognitive engagement required to evaluate it. Seeking a reliable source of medical information to ground a critical decision regarding someone's health need not require the same level of cognitive engagement as searching for a reliable bibliographic source with the goal of writing an essay. However, I submit that the problem is not merely a matter of understanding the degree of engagement required by the domain of the query. Studying epistemic deference in real-word conditions must take into account general constraints that apply to judgments of epistemic evaluation in the context of information search tasks. Fogg's quotation evokes a number of such constraints.

First, there are *time constraints*. In a world in which online content is becoming massively and constantly available, user interactions with the Web naturally tend to become shorter, more frequent and increasingly mediated by search engines. *Information snacking* [21] can be seen as the application to the Web of a known principle of situated cognition that states that organisms tend to externalize the solution of demanding cognitive problems to the environment and use the environment as an external scaffolding to decrease cognitive effort. [22]

The second major class of constraints comes from *epistemic pollution* [23]. The larger the volume of potentially relevant but weakly authoritative information, the more urgent is the need of efficient and cognitively viable skills for source selection. In conditions in which the number of items to evaluate increases beyond control, *a posteriori* evaluative judgments simply become intractable and predictive judgments seem to be the only viable solution. It should be noted, incidentally, that information pollution thrives precisely because of these lightweight, heuristic strategies in which Web users systematically engage. It requires little effort to forge an attractive link luring the user into believing that it will lead to the target source. Fighting epistemic pollution is mainly a matter of detecting cheaters (e.g. sources of unreliable information) on the basis of proximal cues, and this may be an even more effort-consuming task than individually evaluating each source. As Nielsen observes, "information pollution is, [for hungry wolves], like packing the forest with cardboard rabbits" [24]. Good heuristics are those that allow one to tell a fake rabbit from a real one before even starting to hunt it.

The sum of time constraints and constraints imposed by epistemic pollution is the

main rationale in support of the hypothesis on the nature of epistemic reliability judgments that I defend in this paper. I submit that due to these constraints, making a judgment of source reliability on the *World Wide Web* is more likely to be the result of selecting appropriate heuristics, i.e. sufficiently reliable predictive strategies based on link evaluation, than time consuming and cognitively demanding *a posteriori* evaluation processes.

## 2. Heuristics for Epistemic Reliability

I presented in the previous section the main rationale to argue that the study of judgments of epistemic reliability in the context of real information search tasks on the *World Wide Web* should focus on predictive, heuristic strategies rather than in-depth source evaluation processes. Heuristics for the evaluation of credibility of a source as a precondition to epistemic deference can be seen as a subset of a broader class of cognitive heuristics that people assumedly adopt in assessing credibility of electronic information.[25] In this section I will focus on some broad theoretical implications of this hypothesis.

### *2.1. Proximality*

One of the main limitations of traditional studies of Web credibility is the fact that they largely neglected the role of predictive judgments of reliability based on proximal cues about sources of information. The *World Wide Web* is rich with cues that represent (in a more or less reliable way) sources of information. These cues have been referred to in the information foraging literature as the constituents of *information scent*, i.e. a measure of the perceived profitability of a distal source prior to its selection [20]. The hypothesis endorsed by information foraging studies is that information seekers base the choice of optimal navigation patterns on the perceived strength of information scent and on the maximization of scent over effort (e.g. time and length of navigation patterns).

If we accept this assumption, information about a source (e.g. how a source of information is represented in search engine results) becomes much more critical for its evaluation than the actual content it delivers. It is then plausible to assume that the problem facing Web users seeking reliable information is a matter of understanding whether proximal cues (as those afforded by search engines, for instance) are good predictors of target sources. This, I submit, is possible only under the condition that this ecology is stable and sufficiently constrained.

### *2.2. Structured environments*

The *ecology of the Web* has been the object of extensive studies in the information science literature [26-28]. The existence of strong ecological regularities constrains the way in which users learn the structure of the Web and determines to a large extent their preferential strategies in information search behavior. It is plausible to assume that information seekers are situated in this environment and rely on ecological regularities they have learnt in order to select effective solutions to source selection problems.

Cognitive technologies such as search engines aim at improving our information retrieval skills by reducing the cognitive effort required to solve particularly demanding tasks and by increasing the amount of information scent available to the user. In this sense, they tend to favor the selection of simple, effortless and automatic strategies over more costly processes. By enriching the user's ecology with highly informative cues and making this ecology stable, technology aims at reducing information processing requirements on the user.

As Ecological Rationality theories suggest [29,30], stable environments offer ideal conditions to favor the selection of shallow, effortless and relatively rigid computational strategies. These are typical features of modular solutions to the problem of negotiating cognitively demanding problems.

### *2.3. Modularity*

Defenders of the modularity hypothesis insist that modularity arises as a viable solution in stable environments whenever an organism faces a problem of computational tractability of information [31-33]. As Carruthers observes,

> computational processes need to be local–in the sense of having a restricted access to background knowledge in executing their algorithms–if they are to be tractable, avoiding a "computational explosion". And the only known way of realizing this, is to make such processes modular in nature–[31].

If epistemic deference is to be cognitively profitable, then solutions to the problem of estimating the reliability of a source must be computationally tractable. Source evaluation processes whose cost outweighs the benefits of deferring to a source are unlikely to be selected as viable. I will call this a *cognitive affordability constraint* on deferential strategies. The selection of deferees is a paradigmatic case of a problem that has to be solved in a cognitively tractable way by setting limits to background knowledge in order to avoid computational explosion. In the case of reliability judgments, this means finding sufficiently local criteria for estimating the reliability of a source that do not draw in turn on further reliability judgments and so on.

If local inferential strategies can be identified that accurately yield a representation of the trustworthiness of a source, we can then say that the basic conditions are met for the selection of a modular solution to the problem of source evaluation.

## 3. Empirical Research Directions

I have reviewed some of the broad implications of the hypothesis according to which reliability judgments on the Web can in principle be underpinned by highly specialized heuristics based on cues that allow accurate predictions of the reliability of a source. The question that needs to be answered on empirical grounds is then whether – given the ecology of the *World Wide Web* – there are specific heuristics based on information scent that agents systematically use to predict the trustworthiness of a source. In this section, I sketch a program that future research should aim to implement in order to empirically test this hypothesis.

### *3.1. Non-reputational cues in source evaluation*

The first empirical research direction consists in studying how people use *non-reputational* proximal information to decide which sources are worth being selected. I refer to "non-reputational cues" as the class of properties of the proximal representation of a source (e.g. a search engine result) that do not contain explicit information about the credentials of the source. In the case of common search engines, such cues include properties of the title, snippet and URL of an item in a search result page. It is a promising avenue for experimental research to study if we implicitly use properties such as *URL length*, *processing fluency of the snippet* or *density of keywords matching the query* in order to predict whether a specific item in a search engine result page is trustworthy (and hence worth being selected). Experimental research will have to study in particular task-dependent effects, as it is reasonable to expect that users trying to maximize semantic relevance in source selection may not use the same cues as users trying to maximize perceived trustworthiness.

### *3.2. Explicit reputational cues in source evaluation*

A second research direction should focus on the study of the impact of *explicit reputational cues* on judgments of epistemic reliability. The Web offers a plethora of indicators of source “popularity” or “endorsement” that can be regarded as explicit reputational cues. Social software and *Web 2.0* services have already made these indicators a promising avenue for future generations of search engines. [34,35] These cues can be broadly grouped in six different categories:

1. *implicit indicators of individual endorsement* (such as indicators that a specific user selected/visited/purchased an item);
2. *explicit indicators of individual endorsement* (such as explicit ratings produced by specific users);
3. *implicit indicators of socially aggregated endorsement* (such as density of bookmarks or comments per item in social bookmarking systems like *Del.icio.us*, *Digg*, *Reddit* etc.);
4. *explicit indicators of socially aggregated endorsement* (such as average ratings extracted from a user community);
5. *algorithmic endorsement indicators* (such as *PageRank* and similar usage-independent ranking algorithms [36]);
6. *hybrid endorsement indicators* (such as interestingness indicators in *Flickr*, taking into account both explicit user endorsement and usage-independent metrics);

Whereas in general, subjects should have no reason to trust the validity of such reputational cues other than trusting the provider of these cues, it is reasonable to expect that these indicators strongly bias the processes through which we select reliable sources. Experimental research will need to clarify in particular:

- to what extent judgments of epistemic reliability are affected by different types of explicit reputational cues;
- to what extent the overall trust of the subject in the system providing these

cues modulates their judgments.(see for instance the seminal work by Keane and O'Brien, [37]

- to what extent explicit reputational cues override implicit, non-reputational cues.

### *3.3. From reliability heuristics to biases*

Possibly the most interesting question is how these heuristics may result in large-scale biases in deferential behavior, which can be exploited by manipulating the perceived trustworthiness of a source. It has already been shown that the sheer ranking of items in search engine results pages (i.e. the fact that top results attract the vast majority of clicks) produces strong asymmetries in the number of sources that are selected and visited by the majority of users. [38] Similar large-scale asymmetries in the distribution of visits are likely to be found as a result of heuristics that users adopt to evaluate sources of information in a fast and effortless way, especially if the outcome of their selection is fed back to other users. By making the link between individual endorsement and reputational indicators more and more technologically mediated (and hence less transparent to the end user) the Web is already massively biasing the way in which we decide which sources are worth being trusted, selected and visited. Future research will have to clarify the ethical implications of the increasing impenetrability of the reputational cues Web users rely on and consider whether policies need to be introduced to control this phenomenon.

## 4. Conclusions

In this paper, I fleshed out the main rationale, theoretical implications and some potential research directions in the study of processes underlying epistemic reliability judgment on the *World Wide Web*. I proposed that such processes should be regarded as a class of capabilities depending on highly specialized heuristics and that heuristics-based predictive judgments are likely to be more ecologically valid than the kind of evaluative judgments studied so far in the Web credibility literature.

I suggested in particular the conditions under which such heuristics are likely to emerge and stressed how by decreasing the overall cognitive effort involved in source evaluation, they probably are in a better position to describe what users do when engaging in real-world information search behavior.

The rationale for this research program does not rule out the necessity of studying effortful, *a posteriori* evaluative strategies, but calls for a better understanding of the contexts in which these strategies are deployed. Heuristics to ascertain the credibility of sources of information are likely to be privileged only in those cases in which (1) cognitive engagement is low, (2) the ecology in which these strategies apply is sufficiently stable to allow learning, and (3) simple cues are sufficiently accurate to allow the user to cope with epistemic pollution.

I proposed that such conditions may be more common than the current literature on Web credibility would suggest, and that they thus may provide a more realistic account of how we select sources to engage in deferential behavior.

## References


[1] J. M. Joyce. VIII–Epistemic Deference: The Case of Chance. *Proceedings of the Aristotelian Society*, 107(1pt2):187–206, 2007.

[2] F. Recanati. *Oratio Obliqua, Oratio Recta*. MIT Press, Cambridge, MA, 2000.

[3] P. de Brabanter, D. Nicolas, I. Stojanovic, and N. Villanueva Fernández. Deferential utterances. *"Referring to Objects" Interdisciplines Virtual Workshop*, 2005. URL http://www.interdisciplines.org/objects/papers/1.

[4] E. V. Clark. Color, reference, and expertise in language acquisition. *Journal of Experimental Child Psychology*, 94(4):339–343, 2006,

[5] B. J. Fogg and H. Tseng. The elements of computer credibility. *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, New York, 80–87, 1999.

[6] S. Y. Rieh and D. R. Danielson. Credibility: A multidisciplinary framework. In B. Cronin, editor, *Annual Review of Information Science and Technology*, Information Today, volume 41, 307–364, 2007.

[7] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *DUX '03: Proceedings of the 2003 conference on Designing for user experiences*, ACM Press, 15, 2003.

[8] R. Hogarth. Judgment and choice. *The psychology of decision*. John Wiley & Sons, Inc., New York, 1987.

[9] A. H. Eagly and S. Chaiken. *The psychology of attitudes*. Harcourt Brace Jovanovich., New York, 1993.

[10] R. E. Petty and J. Cacioppo. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19:123–205, 1986.

[11] R. E. Petty and D. T. Wegener. The elaboration likelihood model: Current status and controversies. In S. Chaiken and Y. Trope, editors, *Dual-Process Theories in Social Psychology*. Guilford Press, New York, 1999.

[12] A. H. Eagly and S. Chaiken. Attitude strength, attitude structure, and resistance to change. In Petty and Krosnick [23.39].

[13] R. E. Petty, C. Haugtvedt, and S. M. Smith. Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In Petty and Krosnick [23.39].

[14] B. J. Fogg. Prominence interpretation theory: explaining how people assess credibility online. In *CHI '03: CHI '03 extended abstracts on Human factors in computing systems*, ACM Press, New York, 722–723, 2003.

[15] N. C. Wathen and J. Burkell. Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, 53(2):134–144, 2002.

[16] S. Y. Rieh. Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, 53(2): 145–161, 2002.

[17] P. Pirolli. Rational analyses of information foraging on the web. *Cognitive Science*, 29:343–373, 2005.

[18] P. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information* (Oxford Series in HumanTechnology Interaction). Oxford University Press, 2007.

[19] P. Pirolli and S. Card. *The evolutionary ecology of information foraging*, 1997. URL http://citeseer.ist.psu.edu/pirolli97evolutionary.html.

[20] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106(4): 643–675, 1999.

[21] J. Nielsen. *Alertbox: When Search Engines Become Answer Engines*. Website, August 16 2004. URL http://www.useit.com/alertbox/20040816.html.

[22] A. Clark. *Natural born cyborgs: Minds, technologies and the future of human intelligence*. Oxford: Oxford University Press, 2003.

[23] K. Sterelny. Cognitive load and human decision, or, three ways of rolling the rock up hill. In P. Carruthers, S. Laurence, and S. Stich, editors, *The Innate Mind: Culture and Cognition*. Cambridge University Press, Cambridge, 2006.

[24] J. Nielsen. *Alertbox: Information Pollution*. Website, August 11 2003. URL http://www.useit.com/alertbox/20030811.html.

[25] S. S. Sundar. Technology and credibility: Cognitive heuristics cued by modality, agency, interactivity and navigability. In M. J. Metzger and A. J. Flanagin, editors, *Digital Media, Youth, and Credibility*. The MIT Press, 2007.

[26] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the worldwide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[27] B. A. Huberman. *The Laws of the Web : Patterns in the Ecology of Information*. The MIT Press, 2003.

[28] J. E. Pitkow and M. Recker. Results from the First WorldWide Web User Survey. *Computer Networks and ISDN Systems*, 27(2):243–254, 1994.

[29] S. Bullock and P. M. Todd. Made to measure: Ecological rationality in structured environments. *Minds and Machines*, 9(4):497–541, 1999.
[30] P. M. Todd. Fast and frugal heuristics for environmentally bounded minds. In G. Gigerenzer and R. Selten, editors, *Bounded Rationality*. MIT Press, Cambridge, MA, 1999.
[31] P. Carruthers. Moderately massive modularity. In A. O'Hear, editor, *Mind and Persons*. Cambridge University Press, Cambridge, 2003. URL http://www.philosophy.umd.edu/Faculty/pcarruthers/Moderatemodularity.htm.
[32] P. Carruthers. Simple heuristics meet massive modularity. In S. S. P. Carruthers, S. Laurence, editors, *The Innate Mind*. Oxford University Press, Oxford, 2006.
[33] D. Sperber. In defense of massive modularity. In E. Dupoux, editor, *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler*. MIT Press, Cambridge, MA, 2002. URL http://www.dan.sperber.com/modularity.htm.
[34] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, ACM Press, New York, 501–510, 2007.
[35] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the Web? In *JCDL '07: Proceedings of the 2007 conference on digital libraries*, ACM Press, New York, 107–116, 2007
[36] S. Brin and L. Page. The anatomy of a search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Elsevier, Amsterdam, 107-117, 1998.
[37] M. O'Brien and M. T. Keane. Modeling result-list searching in the world wide web: The role of relevance topologies and trust bias. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society,* 2006.
[38] J. Cho and S. Roy. Impact of search engines on page popularity. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, ACM Press, New York, 20–29, 2004.
[39] R. E. Petty and J. A. Krosnick, editors. *Attitude strength: Antecedents and consequences.* Erlbaum, Mahwah, NJ, 1995.

*Current Issues in Computing and Philosophy*
*A. Briggle et al. (Eds.)*
*IOS Press, 2008*



# Author Index

Allo, P. 79
Arkin, R.C. 45
Asaro, P.M. 50
Barnes, T. 145
Brey, P. 91
Bringsjord, S. 156
Casacuberta, D. 103
Croy, M. 145
Fischer, B. 133
Ishii, K. 35
Lanzenberger, M. 184
Li, J. 156
McKinlay, S. 122
Müller, V.C. 116
Noorman, M. 65
Piwek, L. 24
Pohl, M. 184
Rosas, O. 13
Spence, E.H. 3
Stamper, J. 145
Taraborelli, D. 194
Turilli, M. 171
Vallverdú, J. 103
Weiller, D. 133

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank